PR #23890 完整报告

sgl-project/sglang

[spec decoding] add extra attribute 'spec_hidden_size'

合并时间：2026-04-29 10:54

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23890

执行摘要

- 一句话：为 Eagle 推测解码引入 spec_hidden_size，修复 hidden size 不匹配
- 推荐动作：推荐合并并部署。该 PR 改动量小但影响面广，修复了关键的维度不匹配问题，且经过多场景 CI 验证。值得关注的设计决策是将 hc_mult 从模型配置中显式读取，而不是硬编码扩展系数，保持了灵活性和可扩展性。

功能与动机

现有代码在 speculative decoding 流程中始终使用 model_config.hidden_size，但对于某些模型（例如需要 hc_mult 扩展隐藏层维度的架构），draft 模型的隐藏层大小可能不等于 target 模型的 hidden_size，导致张量维度不匹配或显存浪费。PR 通过新增 spec_hidden_size 属性，并统一替换所有相关位置的 hidden_size 引用，修复了此问题。

实现拆解

1. 数据契约扩充- 在 python/sglang/srt/configs/model_config.py 的 _derive_model_shapes 方法中，从 hf_text_config 读取 hc_mult 字段（默认 1），计算出 spec_hidden_size = hidden_size * hc_mult if hc_mult > 1 else hidden_size。
2. Eagle 预处理器替换- 在 eagle_worker.py 的 _draft_preprocess_idle 和 forward_draft_extend_after_decode 中，创建 EagleDraftInput.create_idle_input 时改用 model_config.spec_hidden_size。
3. Verify 流程修正- 在 eagle_info.py 的 verify 方法中，两处创建 EagleDraftInput.create_idle_input 的位置（空闲分支和非空闲分支）均更新为 batch.model_config.spec_hidden_size。
4. Worker 与 Runner 同步- 在 eagle_worker_v2.py、multi_layer_eagle_worker_v2.py、eagle_draft_cuda_graph_runner.py、eagle_draft_extend_cuda_graph_runner.py 中，将对应的 hidden_size 赋值替换为 spec_hidden_size。
5. 调度器 Disaggregation 适配- 在 scheduler.py 的 init_disaggregation 方法中，创建 MetadataBuffers 时，Eagle 模式下的 hidden_size 参数同样改用 model_config.spec_hidden_size。
6. 测试验证- 通过手动触发 1-gpu-5090 与 1-gpu-h100 上的 Eagle 系列测试，CI 全部通过，表明该修正确保了功能正确性。

关键文件：

- python/sglang/srt/configs/model_config.py（模块 模型配置；类别 source；类型 data-contract；符号 _derive_model_shapes, spec_hidden_size）：新增 spec_hidden_size 属性的定义，是本次数据契约变更的源头。
- python/sglang/srt/speculative/eagle_worker.py（模块 推测解码；类别 source；类型 core-logic；符号 _draft_preprocess_idle, forward_draft_extend_after_decode）：该文件是 Eagle 推测解码的核心 worker，修改了两处创建 idle input 的 hidden_size 源。
- python/sglang/srt/managers/scheduler.py（模块 调度器；类别 source；类型 core-logic；符号 init_disaggregation）：调度器在初始化 disaggregation 时使用 spec_hidden_size 来正确设置 metadata 缓冲区大小。
- python/sglang/srt/speculative/eagle_info.py（模块 推测解码；类别 source；类型 core-logic；符号 verify）：verify 方法中两处 create_idle_input 使用了 spec_hidden_size，贯穿验证流程。
- python/sglang/srt/speculative/eagle_draft_extend_cuda_graph_runner.py（模块 推测解码；类别 source；类型 core-logic；符号 init）：CUDA graph runner 在分配 hidden_states 缓冲区时使用 spec_hidden_size，确保图捕获尺寸正确。
- python/sglang/srt/speculative/eagle_worker_v2.py（模块 推测解码；类别 source；类型 core-logic；符号 forward_batch_generation）：Eagle v2 worker 中创建 idle input 时替换 hidden_size 为 spec_hidden_size。
- python/sglang/srt/speculative/multi_layer_eagle_worker_v2.py（模块 推测解码；类别 source；类型 core-logic）：与 eagle_worker_v2 类似的修改，保证多层 Eagle 场景一致性。
- python/sglang/srt/speculative/eagle_draft_cuda_graph_runner.py（模块 推测解码；类别 source；类型 core-logic）：另一个 cuda graph runner 的尺寸调整。

关键符号：_derive_model_shapes, _draft_preprocess_idle, forward_draft_extend_after_decode, verify, forward_batch_generation, init, init_disaggregation

关键源码片段

python/sglang/srt/configs/model_config.py

新增 spec_hidden_size 属性的定义，是本次数据契约变更的源头。

```
# 在 _derive_model_shapes 方法中，设置完 hidden_size 后立即计算 spec_hidden_size
self.hidden_size = self.hf_text_config.hidden_size
# 读取 hc_mult，默认 1（即大多数模型不扩展）
hc_mult = getattr(self.hf_text_config, "hc_mult", 1)
# 当 hc_mult > 1 时使用扩展后的尺寸，否则等于 hidden_size
self.spec_hidden_size = (
    self.hidden_size * hc_mult if hc_mult > 1 else self.hidden_size
)
```

python/sglang/srt/speculative/eagle_worker.py

该文件是 Eagle 推测解码的核心 worker，修改了两处创建 idle input 的 hidden_size 源。

```
# _draft_preprocess_idle 中的改动
batch.spec_info = EagleDraftInput.create_idle_input(
    device=self.device,
    hidden_size=self.model_config.spec_hidden_size, # 之前是 hidden_size
    dtype=self.model_config.dtype,
    topk=self.topk,
    capture_hidden_mode=CaptureHiddenMode.LAST,
)

# forward_draft_extend_after_decode 中的条件分支
hidden_size = (
    self.model_config.hidden_size * 3
    if self.speculative_algorithm.is_eagle3() and self.eagle_use_aux_hidden_state
    else self.model_config.spec_hidden_size # 之前是 hidden_size
)
```

评论区精华

无 review 讨论。PR 作者独立完成并合并，未出现设计争论或悬而未决的问题。

- 暂无高价值评论线程

风险与影响

- 风险：
  1. 向后兼容风险：当 hc_mult 不存在或等于 1 时，spec_hidden_size === hidden_size，所有行为与改动前完全一致，因此现有模型不受影响。
  2. hc_mult 配置缺失：若 hf_text_config 中没有 hc_mult 属性，getattr 默认值为 1，安全兜底。
  3. 多 worker 一致性：eagle_worker_v2.py 和 multi_layer_eagle_worker_v2.py 等均做了替换，遗漏的可能性极低。
  4. 测试覆盖：CI 覆盖了主要的 Eagle 测试用例（basic、infer、beta、disaggregation 等），未发现退化。 - 影响：用户侧：对于使用标准模型（hc_mult<=1）的用户无感知；对于使用需要 hc_mult>1 的模型（如某些多层 Eagle 变体）的用户，此前可能因 hidden size 不匹配而崩溃或产生错误结果，本次变更修复了该问题。系统侧：spec_hidden_size 成为 ModelConfig 的稳定属性，未来其他模块（如异构推理）也可引用。团队侧：统一了 speculative 模块中 hidden size 的语义，降低了后续维护复杂度。
- 风险标记：hc_mult 分支，向后兼容，配置健壮性

关联脉络

- 暂无明显关联 PR