PR #22229 完整报告

sgl-project/sglang

fix(pcg,mm): fix zeroing of input_embeds when replay PCG

合并时间：2026-04-07 20:33

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22229

执行摘要

该 PR 修复了多模态模型在 PCG 重放时 input_embeds 缓冲区清零逻辑的一个索引错误，原代码错误地使用了特征维度索引，导致缓冲区可能残留前次推理内容，影响结果准确性。变更极小（仅 1 行代码），风险低，但未添加测试覆盖。

功能与动机

为什么做：在 PCG 重放前，多模态模型的缓冲区清零逻辑存在错误。buffers.input_embeds 的形状为 (num_tokens, num_dim)，原代码 buffers.input_embeds[:, num_tokens:static_num_tokens].zero_() 错误地使用第二维度（特征维度）索引，当 num_dim > static_num_tokens 时，无法正确清零 tokens 维度，可能导致缓冲区保留前次推理内容，引发错误结果。作者测试 Qwen 3.5-27B 模型（形状 (4096, 5120)，static_num_tokens = 4096）确认了该问题。

实现拆解

修改文件：python/sglang/srt/model_executor/piecewise_cuda_graph_runner.py

关键变更：在 replay_prepare 函数中，将多模态模型 input_embeds 缓冲区的清零逻辑从：

```
buffers.input_embeds[:, num_tokens:static_num_tokens].zero_() # 错误：使用第二维度索引
```

改为：

```
buffers.input_embeds[num_tokens:static_num_tokens].zero_() #
正确：使用第一维度（tokens）索引
```

这确保了缓冲区在 tokens 维度上正确清零，避免残留数据。

评论区精华

review 讨论非常简短：

- Oasis-Git批准并评论："Thanks for your contribution. It is correct."
- 没有其他争议或深入讨论，变更被直接认可为正确修复。

风险与影响

风险：

- 变更极小且逻辑清晰，回归风险低。
- 主要风险是缺少单元测试覆盖，未来可能再次引入类似错误。

影响：

- 对用户：可能提高多模态模型推理准确性，但 PCG 多模态模型默认未启用，影响范围有限。
- 对系统：确保 PCG 重放时缓冲区状态正确，提升可靠性。
- 对团队：作为小型 bugfix，维护了代码质量，但测试缺失需关注。

关联脉络

与近期 PR 的关联：

- PR 22184：同样涉及缓冲区或对象状态管理的 bugfix（如 GenerateReqInput 缓存），关注同步和一致性。
- PR 21952（Gemma 4）和 PR 22073（Qwen3-asr）：涉及多模态模型支持，可能共享类似缓冲区管理逻辑，显示团队在多模态领域的持续扩展。

整体上，该PR是sglang多模态功能演进中的一个小型修复，反映了对缓冲区细节处理的重视。