PR #7094 完整报告

PaddlePaddle/FastDeploy

fix cuda graph capture failure in CI test

合并时间：2026-03-31 11:05

原文链接：http://prhub.com.cn/PaddlePaddle/FastDeploy/pull/7094

执行摘要

本 PR 修复了 FastDeploy 中 speculate decoding 模块在 CI 测试中的 CUDA 图捕获失败问题，通过替换虚拟运行时的 EOS token 为安全值，确保图捕获过程稳定，提升 CI 可靠性。

功能与动机

修复 CI 测试中 CUDA 图捕获失败，原因是在 accept_all=true 的虚拟运行模式下，盲目接受 EOS token 会导致 stop_flags 设置，进而引起后续 MTP 解码步骤的 token 数量不匹配。PR body 明确指出："Previously, accept_all would blindly accept any draft token including EOS, which caused stop_flags to be set in unified_update_model_status, leading to CUDA graph capture failures due to token count mismatch in subsequent MTP decode steps。"

实现拆解

仅修改了 custom_ops/gpu_ops/speculate_decoding/verify_draft_tokens.cu 文件。关键改动点：

- 在 verify_draft_tokens 内核的 accept_all 分支中，添加 EOS 检查逻辑：使用 is_in_end 函数判断 token 是否为 EOS，如果是则替换为 token ID 5。
- 代码示例：cpp if (accept_all) { int64_t token = ctx.step_input_ids_now[i + 1]; if (is_in_end(token, end_tokens, end_length)) { token = 5; } if (ctx.emit_token(i, token)) break; continue; }

评论区精华

Review 中没有技术讨论，只有 reviewer 'freeliuzc' 的批准。代码覆盖率报告显示修改行被测试覆盖，但缺少基准提交报告，这可能影响覆盖率评估的准确性。

风险与影响

- 风险：硬编码 token ID 5 可能在某些模型中不安全；accept_all 模式若被误用于非图捕获场景可能引入问题。但风险较低，因为这是内部调试路径。
- 影响：直接影响 CI 测试稳定性，间接提升团队对 GPU 优化的信心；对用户无直接影响。

关联脉络

与历史 PR 关联：

- PR #7069：修复 CUDA 图捕获的 MoE bug，显示跨模块的图捕获问题共性。
- PR #6738：补充 MTP 测试，反映了 CI 测试覆盖的增强趋势，与本 PR 共同推动测试可靠性。