# PR #7086 完整报告

PaddlePaddle/FastDeploy

[RL][Qwen3VL] Add clear_grpah_opt_backend method to Qwen3VLForConditional···

合并时间：2026-03-31 13:48

原文链接：http://prhub.com.cn/PaddlePaddle/FastDeploy/pull/7086

## 执行摘要

该 PR 为 Qwen3VL 模型新增了 clear_grpah_opt_backend 方法，以支持统一清理 CUDA Graph 缓存，解决了接口缺失问题，影响范围仅限于模型接口扩展。

## 功能与动机

动机源于 Qwen3VLForConditionalGeneration 模型缺少 clear_grpah_opt_backend 方法，导致上层无法通过统一接口清理 CUDA Graph 缓存。PR body 中明确指出："导致上层无法通过统一接口清理 Qwen3VL 模型的 CUDA Graph 缓存"，这确保了模型接口的完整性。

## 实现拆解

实现非常简单，仅修改了 fastdeploy/model_executor/models/qwen3_vl/qwen3_vl.py 文件。在 Qwen3VLForConditionalGeneration 类中新增了以下方法：

```
def clear_grpah_opt_backend(self):
    """Clear graph optimization backend, the captured cuda graph will be cleaned"""
    self.model.clear_grpah_opt_backend(fd_config=self.fd_config)
```

该方法将调用委托给底层 self.model，保持了与其他模型接口的一致性。

## 评论区精华

Review 讨论极少，审核者 CSWYF3634076 仅评论 "LGTM" 并批准，没有其他争议或深入讨论，表明变更被接受为必要补充。

## 风险与影响

风险包括：1) 方法名可能存在拼写错误（'grpah' 可能应为 'graph'），可能影响调用一致性；2) 缺乏单元测试，Codecov 报告显示覆盖率不足；3) 委托调用假设底层模型已实现该方法。影响范围小，仅扩展接口，不改变核心逻辑。

## 关联脉络

从近期历史 PR 看，PR #7094 和 #7069 都涉及 CUDA Graph 优化和缓存管理，表明团队正在持续改进 CUDA Graph 相关功能。本 PR 是这一趋势的一部分，完善了 Qwen3VL 模型的接口。