PR #6998 完整报告

PaddlePaddle/FastDeploy

[Optimization]Streaming requests return complete special tokens.

合并时间：2026-03-26 09:49

原文链接：http://prhub.com.cn/PaddlePaddle/FastDeploy/pull/6998

执行摘要

本 PR 优化了 FastDeploy 中 OpenAI 兼容的 streaming 请求处理，确保在引擎输出被跳过（skipped）时，如果客户端开启 return_token_ids，仍能返回完整的 token ids，从而支持特殊 tokens 的处理。变更涉及核心响应生成逻辑和测试更新，提升 token 流的完整性。

功能与动机

动机源于提升 streaming 场景下 token ids 的完整性。当引擎输出被标记为 skipped 时，默认会跳过该帧，但若客户端开启 return_token_ids，仍需返回对应的 token ids 以避免丢失特殊 tokens。review 中 Copilot 指出：" 优化 OpenAI 兼容的 streaming 响应：当引擎输出被标记为 skipped 时，如果客户端开启 return_token_ids，仍然返回对应的 token ids（从而让'特殊 token/ 完整 token 流'在 streaming 场景下不丢失）。" PR body 未填写具体动机，但代码变更和 review 讨论明确了这一目标。

实现拆解

主要改动集中在两个文件：

- fastdeploy/entrypoints/openai/serving_chat.py：修改 chat_completion_stream_generator 函数，添加条件判断 if output["skipped"] and not request.return_token_ids: continue，并调整 delta_message 内容，例如设置 content="" if output["skipped"] else (output["text"] or "") 和 multimodal_content 占位项。
- fastdeploy/entrypoints/openai/serving_completion.py：类似修改 completion_stream_generator 函数，确保在 skipped 时返回空 text 但保留 token_ids。

关键代码逻辑示例：

```
if output["skipped"] and not request.return_token_ids:
    continue
delta_message = DeltaMessage(
    content="" if output["skipped"] else (output["text"] or ""),
    completion_token_ids=output.get("token_ids") if request.return_token_ids else None
)
```

此外，将 tool_calls 检测提前，以正确影响 finish_reason。测试文件（如 tests/e2e/test_EB_VL_Lite_serving.py）同步更新了 token 计数逻辑，从固定值调整为动态计算，以匹配变更后的行为。

评论区精华

Review 讨论中，Copilot 提出了重要建议：

- 多模态结构问题：" 当启用 multimodal 输出且 output['skipped'] 为 True 时，这里把 delta_message.multimodal_content 设为 [{}] 会导致返回的 multipart item 缺少上游约定的字段 ... 建议在 skipped 时也返回一个结构一致的占位项（例如 type/text 为空字符串）。" 这凸显了设计权衡，需确保客户端兼容性。
- 测试覆盖：" 现有 tests/entrypoints/openai/test_serving_completion.py 的 streaming 用例只覆盖 skipped=False，建议补充 skipped=True 的场景（分别覆盖 return_token_ids=True/False），以避免后续回归。" 表明测试需跟进以保障代码质量。
- 文档规范："PR 标题建议严格遵循模板要求的 [标签] Title 格式 ... 本 PR 描述仍是模板占位，缺少 Motivation/Modifications/Usage 等关键信息。" 最终，reviewer LiqinruiG 以 "LGTM" 批准，但部分建议未明确解决。

风险与影响

风险：

- 响应格式变更风险：修改了核心 streaming 逻辑，可能导致客户端解析错误，特别是在多模态内容结构不一致时（serving_chat.py 中 skipped 时的 multimodal_content 可能缺少 type 或 text 字段）。
- 测试覆盖风险：新增逻辑未在单元测试中充分覆盖 skipped=True 场景，Codecov 报告显示有 2 行缺失覆盖，可能引入回归。
- 兼容性风险：如果现有客户端依赖 skipped 时完全跳过输出的行为，改变后可能需要调整处理逻辑。 影响：
- 用户影响：使用 streaming 并开启 return_token_ids 的客户端能获得更完整的 token 流，便于调试、监控和处理特殊 tokens，提升用户体验。
- 系统影响：优化了 OpenAI 兼容的响应生成，但需确保所有 endpoint（chat 和 completion）行为一致，避免格式错误导致下游问题。
- 团队影响：代码变更较小，但需关注测试更新和维护，以及潜在的多模态结构问题，建议在后续 PR 中补充测试。

关联脉络

从历史 PR 分析中，近期有其他优化和 bugfix 涉及 streaming 或 OpenAI endpoint（如 PR 6680 优化调度，PR 7042 更新 API server 配置），但未发现直接关联本 PR 的修改文件或功能线。本 PR 专注于 streaming token ids 的完整性，是 OpenAI 兼容性优化的一部分，可能为后续相关特性（如特殊 tokens 处理）奠定基础。关联的 cherry-pick PR（7040、7041）表明此变更已反向移植到 release 分支，影响范围扩展至稳定版本。