PR #21755 完整报告

sgl-project/sglang

[diffusion] UX: replace deprecated ORJSONResponse with orjson_response

合并时间：2026-03-31 21:41

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21755

执行摘要

- 一句话：替换已弃用的 ORJSONResponse 为项目内 orjson_response，确保扩散模块 HTTP 响应的序列化一致性。
- 推荐动作：建议工程师在阅读此 PR 时，重点关注 common_api.py 的 decorator 是否应添加 response_class 以维持性能。此 PR 的其他部分为简单替换，适合快速扫描，无需深入分析。

功能与动机

从 PR 标题和变更内容推断，动机是替换已弃用的 FastAPI ORJSONResponse，以避免 deprecation 警告并保持代码兼容性。虽然没有具体的 issue 引用，但这是常见的维护性更新。

实现拆解

变更涉及三个文件：

1. http_server.py：移除 'from fastapi.responses import ORJSONResponse'，添加 'from sglang.srt.utils.json_response import orjson_response'，并将 vertex_generate 函数中的返回语句从 ORJSONResponse(...) 替换为 orjson_response(...)。
2. common_api.py：类似地更新导入，将 available_models 和 retrieve_model 函数的直接返回调用替换，但移除了这两个端点 decorator 的 response_class=ORJSONResponse 参数，可能导致回退到默认 JSONResponse。
3. weights_api.py：更新导入，并将 update_weights_from_disk 和 get_weights_checksum 函数的返回语句替换。

关键文件：

- python/sglang/multimodal_gen/runtime/entrypoints/http_server.py（模块 diffusion/http_server）：主 HTTP 服务器入口，替换直接响应调用，确保基础功能更新
- python/sglang/multimodal_gen/runtime/entrypoints/openai/common_api.py（模块 diffusion/openai_api）：OpenAI 兼容 API 入口，decorator 变更可能导致性能回退，是关键风险点
- python/sglang/multimodal_gen/runtime/entrypoints/post_training/weights_api.py（模块 diffusion/post_training）：权重更新 API，替换响应调用，属于常规维护

关键符号：vertex_generate, available_models, retrieve_model, update_weights_from_disk, get_weights_checksum

评论区精华

review 中，gemini-code-assist[bot] 指出，common_api.py 中移除 response_class 参数会使得 /models 和 /models/{model:path} 端点回退到默认 JSONResponse，损失 orjson 的性能优势（如 numpy 支持）。建议使用 SGLangORJSONResponse 作为 response_class 来维持优化序列化。但此评论未被回复，PR 已合并，该问题可能未解决。

- Decorator response_class 移除导致的序列化问题 (performance): 建议使用 SGLangORJSONResponse 作为 response_class，但未在 PR 中实施或讨论。

风险与影响

- 风险：主要风险在于 common_api.py 的端点可能失去自定义序列化行为，导致性能下降（序列化速度变慢）和功能性问题（如不支持 numpy 类型）。此外，如果项目依赖特定的序列化选项，可能出现不一致响应。风险文件为 common_api.py。
- 影响：影响范围仅限于扩散模块的 HTTP API 响应序列化，用户可能察觉不到变化或仅有轻微延迟。开发团队需要关注此变更以确保性能一致，但整体影响较小，属于低风险维护。
- 风险标记：潜在性能回退，序列化不一致性

关联脉络

- PR #21746 [diffusion] Fix typo: 同为扩散模块的维护性 PR，共享 'diffusion' 标签，但无直接技术关联