# PR #22174 完整报告

sgl-project/sglang

UX: clean loggings

合并时间：2026-04-08 09:46

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22174

## 执行摘要

- 一句话：修复 FastAPI 弃用警告并统一多模态编码器参数命名，提升日志清晰度。
- 推荐动作：建议快速浏览以了解代码清理实践，重点关注 json_response.py 的重构设计，但整体变更较小，无需深入精读。

## 功能与动机

PR body 中显示了 FastAPIDeprecationWarning: 'ORJSONResponse is deprecated'，动机是消除此警告以改善用户体验和代码维护性，避免日志污染。

## 实现拆解

实现分为三个部分：1) 在 mistral_3.py 和 qwen2_5vl.py 中将参数名从 input_embeds 更正为 inputs_embeds，修复拼写错误；2) 在 logging_utils.py 的 globally_suppress_loggers 函数中添加 'flash_attn.cute.cache_utils' 到抑制列表，减少日志输出；3) 在 json_response.py 中重构 SGLangORJSONResponse 类，使其继承自 Response 而非 ORJSONResponse，并更新 orjson_response 函数以使用新类，移除弃用依赖。

关键文件：

- python/sglang/multimodal_gen/runtime/models/encoders/mistral_3.py（模块 multimodal_gen/encoders）：修正参数名拼写错误，从 input_embeds 改为 inputs_embeds，确保多模态编码器 forward 方法调用正确。
- python/sglang/multimodal_gen/runtime/models/encoders/qwen2_5vl.py（模块 multimodal_gen/encoders）：同上，统一参数命名以避免潜在调用错误。
- python/sglang/multimodal_gen/runtime/utils/logging_utils.py（模块 multimodal_gen/logging）：添加 flash_attn.cute.cache_utils 到全局日志抑制列表，减少无关日志输出，提升日志清晰度。
- python/sglang/srt/utils/json_response.py（模块 srt/utils）：重构 SGLangORJSONResponse 类，解决 FastAPI 弃用警告，优化 JSON 序列化逻辑，是核心变更点。

关键符号：forward (in mistral_3.py and qwen2_5vl.py), globally_suppress_loggers, SGLangORJSONResponse.render, orjson_response

## 评论区精华

review 中仅有一个评论：gemini-code-assist[bot] 指出 SGLangORJSONResponse.render 方法应检查 content 是否为 bytes 以避免 TypeError，并提供代码建议。但此建议未被采纳，最终代码未包含该检查，评论状态为已忽略。

- SGLangORJSONResponse.render 方法安全性 (correctness): 建议未被采纳，最终代码未实现此检查，可能引入潜在序列化错误。

## 风险与影响

- 风险：风险包括：1) 参数名变更在 mistral_3.py 和 qwen2_5vl.py 中可能影响内部调用链，但由于是拼写错误修复，风险较低；2) json_response.py 中 SGLangORJSONResponse.render 方法未处理 bytes 类型 content，可能导致序列化错误，增加潜在 bug；3) 日志抑制添加可能掩盖关键调试信息，但影响限于特定模块。
- 影响：影响范围主要限于多模态生成模块和日志系统：对终端用户无直接影响，但能减少警告日志，提升开发体验；系统层面优化了代码整洁度，降低维护负担；团队需注意 JSON 响应类的变更可能影响序列化逻辑。
- 风险标记：参数名变更风险，JSON 序列化潜在错误

## 关联脉络

- PR #22251 [diffusion] CI: fix consistency check: 同属 diffusion 模块，涉及多模态生成 CI 测试，本 PR 的日志清理可能与其相关。
- PR #22229 fix(pcg,mm): fix zeroing of input_embeds when replay PCG: 涉及多模态模型和 input_embeds 相关修复，与本 PR 的参数名修正有间接关联。