

PR #30621 完整报告

sgl-project/sglang

Fix image URL response for multiple outputs

合并时间: 2026-07-15 19:49

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/30621>

执行摘要

- 一句话: 修复多输出时 URL 只返回单个的问题, 新增 variant 回退支持
- 推荐动作: 此 PR 值得精读, 原因:
 - 清晰的变体索引设计, 通过列表和索引映射处理多输出。
 - 良好的回退逻辑, 兼容旧数据结构。
 - 并发上传优化和明确的错误提示。
 - 测试覆盖全面 (8 个用例)。建议关注 `_build_image_response_kwargs` 中 `url` 分支的循环构建方式, 以及下载端点如何使用辅助函数处理 variant。

功能与动机

Issue #30648 报告当 `n=2` 且 `response_format='url'` 时, 响应只返回一个 `data` 项, 而 `b64_json` 正常返回多个。用户期望每个输出对应一个 URL。PR 旨在使 URL 格式响应与 `b64_json` 格式对齐, 并确保本地持久化输出也能通过 variant 查询参数下载。

实现拆解

1. 新增辅助函数: `_get_request_field_or_extra` 统一获取字段;
`_upload_and_cleanup_images` 使用 `asyncio.gather` 并发上传多个文件;
`_fallback_image_urls` 根据输出数量生成 variant 形式的本地回退 URL 列表;
`_select_image_variant_path`、`_image_variant_index`、`_select_image_variant_cloud_url`、`_raise_if_image_variant_not_found` 支持按 variant 检索文件路径或云 URL。
2. 修改 `_build_image_response_kwargs`: 接受可选的 `cloud_urls` 和 `fallback_urls` 列表参数, 为每个输出文件独立构建 `ImageResponseData`, 而非使用单个 url; 移除必须配置云存储的限制。
3. 端点 `generations` 和 `edits` 适配: 调用 `_upload_and_cleanup_images` 上传所有文件并收集 URL 列表; 在 `IMAGE_STORE` 中存储 `file_paths`、`urls`、`num_outputs`; 传递列表至响应构建函数。
4. 下载端点 `download_image_content` 重构: 使用 `_select_image_variant_path` 读取 indexed file paths; 若文件未持久化但云 URL 存在则返回重定向; 使用 `_raise_if_image_variant_not_found` 给出明确的 404 错误 (如 'Image variant 5 not found')。

5. 测试覆盖：新增单元测试文件，覆盖多输出 URL 响应、variant 回退、路径选择、cloud URL 对齐、越界变体错误等场景。

关键文件：

- python/sglang/multimodal_gen/runtime/entrypoints/openai/image_api.py (模块 图像 API; 类别 source; 类型 entrypoint; 符号 _get_request_field_or_extra, _upload_and_cleanup_images, _fallback_image_urls, _select_image_variant_path) : 核心变更文件，修改响应构建、上传、下载端点以支持多输出 URL 和 variant 回退。
- python/sglang/multimodal_gen/test/unit/test_openai_image_api.py (模块 单元测试; 类别 test; 类型 test-coverage; 符号 test_url_response_returns_one_item_per_output_path, test_url_response_uses_variant_fallback_urls_for_multiple_persistent_outputs, test_select_image_variant_path_reads_indexed_file_paths, test_raise_if_image_variant_not_found_handles_out_of_range_variant) : 新增完整单元测试，验证多输出 URL 响应、variant 路径选择、越界错误等场景，保证修复质量。

关键符号：_build_image_response_kwargs, _fallback_image_urls, _select_image_variant_path, _select_image_variant_cloud_url, _raise_if_image_variant_not_found, _upload_and_cleanup_images, generations, edits, download_image_content

关键源码片段

[python/sglang/multimodal_gen/runtime/entrypoints/openai/image_api.py](#)

核心变更文件，修改响应构建、上传、下载端点以支持多输出 URL 和 variant 回退。

```
def _fallback_image_urls(
    request_id: str, num_outputs: int, is_persistent: bool
) -> list[str] | None:
    '''生成回退 URL 列表。每个输出对应一个 variant URL。'''
    if not is_persistent:
        return None
    if num_outputs <= 1:
        # 单个输出时不加 variant 参数，保持向后兼容
        return [f'/v1/images/{request_id}/content']
    # 多个输出时，每个输出使用不同的 variant 索引
    return [
        f'/v1/images/{request_id}/content?variant={idx}' for idx in range(num_outputs)
    ]
```

_build_image_response_kwargs 中的 url 分支被重写为循环：

```
elif resp_format == 'url':
    data = []
    for idx, path in enumerate(save_file_path_list):
        # 优先使用云存储 URL（可能为 None），否则使用本地回退 URL
        url = (
            cloud_urls[idx]
            if cloud_urls and idx < len(cloud_urls) and cloud_urls[idx]
```

```

        else fallback_urls[idx]
        if fallback_urls and idx < len(fallback_urls)
        else None
    )
    data.append(
        ImageResponseData(
            url=url,
            revised_prompt=prompt,
            file_path=os.path.abspath(path) if is_persistent else None,
        )
    )
ret = {'data': data}

```

评论区精华

Review 中 gemini-code-assist[bot] 提出三个关键改进：

- cloud_urls 索引对齐：原始实现过滤了 None 值导致索引不对齐，建议存储包含 None 的原始列表。已修复。
- 并发上传：初始版本顺序 await 逐个上传，建议使用 asyncio.gather 并发上传以提升响应速度。已采用。
- 下载端点错误提示：当请求的 variant 已上传到云但本地文件不存在时，原先使用 `item.get('url')` 可能指向错误变体，建议使用 `_select_image_variant_cloud_url` 提供准确的 cloud URL。已修改。
 - cloud_urls 列表索引对齐 (correctness): 已修复，存储原始 cloud_urls 列表（含 None），保持索引对应。
 - 并发上传优化 (performance): 已采用 `_upload_and_cleanup_images` 使用 `asyncio.gather`。
 - 下载端点变体错误提示 (correctness): 已修改代码使用 `_select_image_variant_cloud_url` 获取准确 URL。

风险与影响

- 风险：主要风险包括：
 - 数据兼容性：IMAGE_STORE 新增 urls 和 file_paths 列表字段，旧持久化数据只有 url 和 file_path。辅助函数已包含回退逻辑，兼容旧结构。
 - 并发上传：使用 asyncio.gather 增加内存占用，但图像文件通常不大，风险可控。
 - Variant 解析：_image_variant_index 处理了类型转换异常，但超大整数值可能引发性能边缘问题。
 - 依赖：仅新增 import asyncio，无其他依赖。
 - 影响：影响所有使用 OpenAI 图像生成 API (/v1/images/generations 和 /v1/images/edits) 且设置 response_format='url' 及 n>1 的用户。修复后 URL 响应与 b64_json 行为一致，每个输出返回独立 URL。本地持久化模式下可通过 variant 查询参数下载特定输出。后端存储结构轻微扩展，不影响现有数据。团队需验证向后兼容性。

- 风险标记：数据兼容性，并发控制

关联脉络

- PR #30648 [Bug] OpenAI image API returns only one URL when $n > 1$ and `response_format="url"`: 关联 issue，描述相同 bug，当前 PR 修复此问题。
- PR #30735 Fix OpenAI image URL for multi-output (dup): 针对同一问题的重复 PR，当前 PR 更完善，讨论中提及。