

PR #30148 完整报告

sgl-project/sglang

[diffusion] Pass progressive params through image API

合并时间: 2026-07-06 14:45

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/30148>

执行摘要

- 一句话: 为扩散模型图像 API 暴露渐进式分辨率参数
- 推荐动作: 值得关注: 展示了如何在保持向后兼容的前提下扩展 OpenAI 兼容 API。PR 虽小但设计合理: 显式字段 + `extra_body` 回退。建议后续补充单元测试和 API 文档更新。

功能与动机

用户希望在调用 `client.images.generate(...)` 时直接传递渐进式分辨率参数 (`progressive_mode`、`progressive_levels`、`progressive_delta`)，而之前只能通过 `extra_body` 间接传入。该 PR 将这些参数显式声明，使 API 更符合 OpenAI 扩展规范，降低使用门槛。

实现拆解

1. 在 `protocol.py` 中添加字段声明— 在 `ImageGenerationsRequest` 类末尾新增三个可选字段: `progressive_mode: Optional[str]`、`progressive_levels: Optional[int]`、`progressive_delta: Optional[float]`，并添加注释说明用途。
2. 在 `image_api.py` 中透传参数— 在构造 `SamplingParams` 的字典中添加这三个参数，优先取请求对象中的显式值，None 时回退到 `_get_extra_field(request, ...)` 从 `extra_body` 中获取，保证向后兼容。
3. 无测试配套变更— 仅对两个源文件进行了增量改动，未添加专用测试用例。

关键文件:

- `python/sglang/multimodal_gen/runtime/entrypoints/openai/protocol.py` (模块 API 协议层; 类别 source; 类型 core-logic; 符号 `ImageGenerationsRequest`): 定义 `ImageGenerationsRequest` 数据模型，新增三个渐进式字段，是 API 协议层的核心变更。
- `python/sglang/multimodal_gen/runtime/entrypoints/openai/image_api.py` (模块 图像 API 入口; 类别 source; 类型 endpoint; 符号 `generations`): 入口函数 `generations` 中将新增字段从 `request` 对象映射到 `SamplingParams`，是实际生效的透传点。

关键符号: `generations`, `_get_extra_field`

关键源码片段

`python/sglang/multimodal_gen/runtime/entrypoints/openai/protocol.py`

定义 `ImageGenerationsRequest` 数据模型，新增三个渐进式字段，是 API 协议层的核心变更。

```
# python/sglang/multimodal_gen/runtime/entrypoints/openai/protocol.py
class ImageGenerationsRequest(BaseModel):
    # ... 已有字段 ...
    diffusers_kwargs: Optional[Dict[str, Any]] = None # kwargs for diffusers backend
    # Performance profiling
    perf_dump_path: Optional[str] = None
    # 新增的渐进式分辨率生成参数，之前只能通过 extra_body 传入
    progressive_mode: Optional[str] = None # e.g., "dct_rewind"
    progressive_levels: Optional[int] = None # e.g., 1
    progressive_delta: Optional[float] = None # e.g., 0.01
```

`python/sglang/multimodal_gen/runtime/entrypoints/openai/image_api.py`

入口函数 `generations` 中将新增字段从 `request` 对象映射到 `SamplingParams`，是实际生效的透传点。

```
# python/sglang/multimodal_gen/runtime/entrypoints/openai/image_api.py
# 在 SamplingParams 构造字典中添加渐进式分辨率参数
sampling_params = SamplingParams(
    # ... 已有参数 ...
    preset=_get_extra_field(request, "preset"),
    # 新增三个字段：优先取请求对象中的显式值，None 时从 extra_body 回退
    progressive_mode=(
        request.progressive_mode
        if request.progressive_mode is not None
        else _get_extra_field(request, "progressive_mode")
    ),
    progressive_levels=(
        request.progressive_levels
        if request.progressive_levels is not None
        else _get_extra_field(request, "progressive_levels")
    ),
    progressive_delta=(
        request.progressive_delta
        if request.progressive_delta is not None
        else _get_extra_field(request, "progressive_delta")
    ),
)
```

评论区精华

该 PR 的 review 过程简洁，无深入讨论。维护者 mickqian 触发 CI 并最终批准合并，PR 作者在 CI 通过后请求合并，过程顺利。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。三个字段均为 Optional 且默认 None，现有请求不受影响。参数仅做透传，不会改变模型计算逻辑。_get_extra_field 回退机制确保从 extra_body 传入的旧方式继续兼容。但缺少单元测试覆盖新字段的序列化和回退逻辑，若后续重构 SamplingParams 或 _get_extra_field 行为，可能遗漏此处的依赖。
- 影响：对用户：使用扩散模型的用户现在可以直接通过 API 参数指定渐进式分辨率生成，无需再拼凑 extra_body。对系统：无性能或功能影响，仅入口层多传递三个可能为 None 的字段。对团队：降低了 extra_body 的滥用，利于 API 文档化和后续维护。
- 风险标记：缺少测试覆盖

关联脉络

- PR #23049 [Diffusion] Diffusion model support log-requests: 同为 diffusion 模块的 OpenAI API 扩展，涉及协议和入口的类似改动模式。