PR #26926 完整报告

sgl-project/sglang

[diffusion] feat: improve cosmos3 serve API support

合并时间：2026-06-02 00:53

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26926

执行摘要

- 一句话：改进 Cosmos3 服务 API，支持 vLLM-Omni 风格参数与同步视频端点
- 推荐动作：该 PR 值得精读，尤其设计决策如使用 model_extra 传递额外参数以保持协议稳定性、同步视频端点的轮询实现、以及 guardrail 逐请求控制。建议关注 review 中未解决的问题（flow_shift 安全访问、资源泄漏）是否在后续修复。

功能与动机

本 PR 增强 SGLang 对 NVIDIA Cosmos3 模型（T2I、T2V、I2V）的服务能力，使其更符合 OpenAI API 规范（vLLM-Omni 风格）。主要驱动来自社区对阻塞式视频生成和细粒度参数控制的需求。同时处理不支持的生成模式（声音、动作控制、V2V），明确返回 400 错误，并注册公开 checkpoint 以便用户直接使用。

实现拆解

1. 协议与参数扩展（protocol.py、sampling_params.py、schedule_batch.py、server_args.py）：在请求模型中添加 max_sequence_length、flow_shift 等字段；通过 model_extra 保留 use_duration_template、use_system_prompt 等不稳定参数；在调度批次中传递这些参数。
2. 视频 API 改造（video_api.py）：新增辅助函数 extra_value、_parse_form_extra_value 用于提取 extra 字段；新增 _reject_unsupported_cosmos3_modes 校验，对 generate_sound、action_mode、V2V 条件直接 400 拒绝；新增 /v1/videos/sync 同步视频端点，采用轮询方式阻塞返回 MP4；在 build_sampling_params 中传入 max_sequence_length、flow_shift、use* 等参数。
3. 图像 API 改造（image_api.py）：重写 _get_extra_field 支持从多个容器（extra_body、extra_json、extra_args、extra_params）回退查找；新增 _parse_extra_container 统一解析嵌套 JSON；对 Cosmos3 模型，默认输出格式从 jpg 改为 png，并传入额外参数。
4. 公共工具函数（utils.py）：新增 flatten_extra_params 函数，将 vLLM-Omni 风格的 extra_params 展开为平级字段，并处理 guardrails 字段映射。
5. Guardrail 控制改进（cosmos3_guardrails.py）：新增 is_cosmos_guardrail_available 函数检查包是否安装并缓存结果；在 forward 中根据 batch.use_guardrails 跳过检查。
6. Cosmos3 流水线消费新参数（cosmos3.py、cosmos3_pipeline.py）：TimestepPreparationStage 支持 flow_shift 动态设置 scheduler 位移；

use_duration_template 与 use_system_prompt 从 batch 回退到 pipeline_config 默认值；后处理时根据 batch.use_guardrails 控制视频安全检测。

7. 测试覆盖（test_cosmos3.py）：新增 TestCosmos3ModelResolution 验证公开 checkpoint 正确注册；TestCosmos3OpenAIProtocol 确保新字段在模型字段中而额外字段不在；TestCosmos3Guardrails 测试 guardrail 可用性。
8. 文档（Cosmos3.mdx）：详尽的使用指南，列出所有支持 checkpoint、参数说明、示例 curl 命令。

关键文件：

- python/sglang/multimodal_gen/runtime/entrypoints/openai/video_api.py（模块 视频入口；类别 source；类型 entrypoint；符号 _extra_value, _parse_form_extra_value, _reject_unsupported_cosmos3_modes, form_value）：核心变更文件：添加 Cosmos3 专用参数提取、同步视频端点、不支持的生成模式拒绝逻辑。
- python/sglang/multimodal_gen/runtime/entrypoints/openai/image_api.py（模块 图像入口；类别 source；类型 entrypoint；符号 _parse_extra_container）：图像生成端点扩展，支持额外容器查找参数，Cosmos3 默认使用 PNG 格式，传递新参数。
- python/sglang/multimodal_gen/test/unit/test_cosmos3.py（模块 Cosmos3 测试；类别 test；类型 test-coverage；符号 TestCosmos3ModelResolution, test_hf_checkpoint_uses_registered_native_pipeline_config, TestCosmos3OpenAIProtocol, test_cosmos3_private_fields_are_extra_fields）：新增单元测试覆盖 checkpoint 注册、协议字段验证、不支持的生成模式测试及 guardrail 可用性。
- python/sglang/multimodal_gen/runtime/entrypoints/openai/utils.py（模块 公共工具；类别 source；类型 dependency-wiring；符号 flatten_extra_params）：新增 flatten_extra_params 工具函数，用于将 vLLM-Omni 风格的嵌套额外参数展开为平级字段，并处理 guardrails 到 use_guardrails 的映射。
- python/sglang/multimodal_gen/runtime/pipelines_core/stages/model_specific_stages/cosmos3_guardrails.py（模块 防护机制；类别 source；类型 data-contract；符号 is_cosmos_guardrail_available）：新增 is_cosmos_guardrail_available 函数缓存检查 cosmos_guardrail 包是否安装；forward 方法中根据 batch.use_guardrails 跳过安全检查。
- python/sglang/multimodal_gen/runtime/pipelines_core/stages/model_specific_stages/cosmos3.py（模块 流水线核心；类别 source；类型 data-contract）：Cosmos3 扩散流水线核心 stages，消费新增参数：flow_shift 动态设置 scheduler、use_duration_template/use_system_prompt 从 batch 回退到 config、后处理时根据 use_guardrails 控制安全检测。
- python/sglang/multimodal_gen/runtime/pipelines_core/schedule_batch.py（模块 批次调度；类别 source；类型 core-logic）：调度批次类增加 use_guardrails、flow_shift 等属性，并安全访问 pipeline_config.flow_shift（review 指出风险）。
- python/sglang/multimodal_gen/runtime/server_args.py（模块 服务器参数；类别 source；类型 core-logic）：服务器参数配置中添加 Cosmos3 相关字段（如 flow_shift、use_duration_template 等）的默认值支持。
- python/sglang/multimodal_gen/runtime/entrypoints/openai/protocol.py（模块 请求协议；类别 source；类型 core-logic）：请求协议模型中添加 max_sequence_length、

flow_shift 等公开字段，维持后端稳定性。

- python/sglang/multimodal_gen/configs/sample/sampling_params.py（模块 采样参数；类别 source；类型 core-logic）：采样参数类新增 max_sequence_length、flow_shift 等字段，保证参数传递完整。
- docs_new/cookbook/diffusion/Cosmos/Cosmos3.mdx（模块 文档；类别 other；类型 documentation）：全新文档，详尽说明 Cosmos3 模型使用方式、支持参数、示例命令，方便用户上手。

关键符号：_extra_value, _parse_form_extra_value, _reject_unsupported_cosmos3_modes, _get_extra_field, _parse_extra_container, flatten_extra_params, is_cosmos_guardrail_available, test_cosmos3_private_fields_are_extra_fields, test_unsupported_cosmos3_modes_allow_falsy_extra_fields, test_hf_checkpoint_uses_registered_native_pipeline_config

关键源码片段

python/sglang/multimodal_gen/runtime/entrypoints/openai/video_api.py

核心变更文件：添加 Cosmos3 专用参数提取、同步视频端点、不支持的生成模式拒绝逻辑。

```
# python/sglang/multimodal_gen/runtime/entrypoints/openai/video_api.py

def _extra_value(request: VideoGenerationsRequest, name: str) -> Any:
    """从请求的 model_extra 字典中安全获取额外参数。"""
    return (request.model_extra or {}).get(name)


def _parse_form_extra_value(value: Any) -> Any:
    """解析表单传入的额外值，如果它是 JSON 则反序列化。"""
    if not isinstance(value, str):
        return value
    try:
        return json.loads(value)
    except Exception:
        return value


def _reject_unsupported_cosmos3_modes(
    req: VideoGenerationsRequest, model_path: str | None
) -> None:
    """如果当前模型是 Cosmos3，检查并拒绝不支持的生成模式。"""
    if "cosmos3" not in (model_path or "").lower():
        return

    extra = req.model_extra or {}
    if extra.get("generate_sound"):
        raise HTTPException(
```

```
            status_code=400,
            detail="Cosmos3 video-with-sound is not supported by SGLang yet; omit generate_
            sound for video-only generation.",
        )
    if extra.get("action_mode"):
        raise HTTPException(
            status_code=400,
            detail="Cosmos3 action generation is not supported by SGLang yet.",
        )
    if extra.get("condition_frame_indexes_vision") or extra.get("condition_video_keep"):
        raise HTTPException(
            status_code=400,
            detail="Cosmos3 video-to-video conditioning is not supported by SGLang yet.",
        )

# 在 _build_video_sampling_params 中新增 Cosmos3 专用参数传递：
# max_sequence_length=request.max_sequence_length,
# flow_shift=request.flow_shift,
# use_duration_template=_extra_value(request, "use_duration_template"),
# use_resolution_template=_extra_value(request, "use_resolution_template"),
# use_system_prompt=_extra_value(request, "use_system_prompt"),
# use_guardrails=_extra_value(request, "use_guardrails"),
```

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

图像生成端点扩展，支持额外容器查找参数，Cosmos3 默认使用 PNG 格式，传递新参数。

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

def _get_extra_field(request, field_name):
    """增强版：先查 model_extra，再查 guardrails 别名，最后遍历多个容器。"""
    extra = request.model_extra or {}
    value = extra.get(field_name)
    if value is not None:
        return value
    if field_name == "use_guardrails" and extra.get("guardrails") is not None:
        return extra["guardrails"]

    for container_name in ("extra_body", "extra_json", "extra_args", "extra_params"):
        value = _parse_extra_container(extra.get(container_name)).get(field_name)
        if value is not None:
            return value

    return value


def _parse_extra_container(value: Any) -> dict[str, Any]:
    """解析容器字段，支持 JSON 字符串或字典，并应用 flatten_extra_params。"""
    if isinstance(value, str):
        try:
```

```
            value = json.loads(value)
        except Exception:
            return {}
    if isinstance(value, dict):
        return flatten_extra_params(dict(value))
    return {}

# 在 generations 端点中，识别 Cosmos3 模型并调整行为：
# is_cosmos3 = "cosmos3" in (server_args.model_path or "").lower()
# ext = "png" if is_cosmos3 and request.output_format is None else ...
# 并传递 max_sequence_length、flow_shift 等参数到 build_sampling_params
```

python/sglang/multimodal_gen/test/unit/test_cosmos3.py

新增单元测试覆盖 checkpoint 注册、协议字段验证、不支持的生成模式测试及 guardrail 可用性。

```
# python/sglang/multimodal_gen/test/unit/test_cosmos3.py

class TestCosmos3OpenAIProtocol(unittest.TestCase):
    """Verify Cosmos3-only knobs stay out of the stable request schema."""

    def test_cosmos3_private_fields_are_extra_fields(self):
        # 稳定的字段出现在模型字段中，不稳定的参数只出现在 extra 中
        for request_cls in (ImageGenerationsRequest, VideoGenerationsRequest):
            with self.subTest(request_cls=request_cls.__name__):
                self.assertIn("max_sequence_length", request_cls.model_fields)
                self.assertIn("flow_shift", request_cls.model_fields)
                self.assertNotIn("use_duration_template", request_cls.model_fields)
                self.assertNotIn("use_resolution_template", request_cls.model_fields)
                self.assertNotIn("use_system_prompt", request_cls.model_fields)
                self.assertNotIn("use_guardrails", request_cls.model_fields)
                self.assertNotIn("generate_sound", VideoGenerationsRequest.model_fields)
                self.assertNotIn("sound_duration", VideoGenerationsRequest.model_fields)

    def test_unsupported_cosmos3_modes_allow_falsy_extra_fields(self):
        # 验证 falsy 值不会触发拒绝（review 中建议检查 truthiness）
        req = VideoGenerationsRequest(
            prompt="test",
            model_extra={
                "generate_sound": False,
                "action_mode": "",
                "condition_frame_indexes_vision": [],
            },
        )
        # 应当不抛出 HTTPException
        _reject_unsupported_cosmos3_modes(req, model_path="nvidia/Cosmos3-Super")
```

评论区精华

gemini-code-assist[bot] 在 review 中提出三条建议：

1. 在 schedule_batch.py 中使用 getattr 安全获取 flow_shift，避免其他模型因缺失属性而 AttributeError。
2. 在同步视频生成轮询循环中检查客户端断连，防止资源泄漏。
3. 在 _reject_unsupported_cosmos3_modes 中检查字段 truthiness 而非 is not None，以避免显式设置 falsy 值（如 False、""）时被误拒。 这些建议均以 COMMENTED 状态提出，PR 已合并，但未确认是否全部采纳。

- 安全获取 flow_shift 属性 (correctness): 建议已提出，PR 合并时未确认是否修改；后续需警惕此风险。
- 检查客户端断连避免资源泄漏 (performance): 建议已提出，但合并前未见修改；该同步端点可能存在资源泄漏风险。
- 检查 unsupported 字段的 truthiness (correctness): 建议已提出，但合并前未修改；当前实现仍使用 is not None 可能导致边界问题。

风险与影响

- 风险：
  1. 同步视频端点 /v1/videos/sync 采用轮询实现，若客户端断连未检测，将导致协程泄漏（review 已指出）。
  2. 直接访问 server_args.pipeline_config.flow_shift 可能在其他模型上引发 AttributeError，需改用 getattr。
  3. 请求级 guardrail 控制（use_guardrails）为新增行为，若默认启用，可能影响原有期望无安全检查的用户（可通过全局环境变量禁用）。
  4. 不支持的生成模式检查中，使用 extra.get(field) is not None 判断会拒绝显式设为逻辑假值的参数，破坏前端 disabled 语义。 - 影响：对用户：Cosmos3 用户现在可以使用 OpenAI 兼容参数，同步视频返回 MP4，更易集成。对系统：新增端点增加少量负载，但仅对 Cosmos3 模型有效。对团队：需维护 Cosmos3 特定参数映射和 guardrail 逻辑，但通用工具函数 flatten_extra_params 可被其他模型复用。 - 风险标记：同步端点资源泄漏风险，属性访问潜在 AttributeError，参数 falsy 值误拒风险，默认 guardrail 影响感知

关联脉络

- 暂无明显关联 PR