# PR #26492 完整报告

sgl-project/sglang

[diffusion] model: update to new model format

合并时间：2026-05-29 12:08

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26492

## 执行摘要

- 一句话：更新 Cosmos3 至新 diffusers 格式，使用外部 guardrails 包
- 推荐动作：建议开发者和维护者重点阅读 cosmos3_guardrails.py 的重写思路和配置文件的映射设计。本 PR 展示了将自实现功能迁移到专用外部包的典型模式，以及如何在演进中保持与 upstream checkpoint 的一致。对于部署人员，需注意新增的 pip 依赖。

## 功能与动机

PR #24994 添加了对一个尚未发布的 world model 的初始支持。最终 checkpoint 格式略有变化，本 PR 更新代码库以适配新版 checkpoint。此外，移除了自定义 guardrails 代码，改用包驱动的方式（用户需在运行模型前通过 pip 手动安装）。

## 实现拆解

1. 更新权重参数名映射：在 python/sglang/multimodal_gen/configs/models/dits/cosmos3video.py 的 _build_cosmos3_param_names_mapping 中，移除所有源键的 model. 前缀；将 UND 路径的 q_proj/k_proj/v_proj 改为 to_q/to_k/to_v；GEN 路径的 q_proj_moe_gen 改名为 add_q_proj/add_k_proj/add_v_proj，o_proj_moe_gen 改为 to_add_out，q/k_norm_moe_gen 改为 norm_added_q/k；time_embedder.mlp.0/2 直接改为直通 linear_1/2（因新版 checkpoint 已使用 linear_* 命名）。
2. 重命名潜变量投影层：在 python/sglang/multimodal_gen/runtime/models/dits/cosmos3video.py 中，将 vae2llm 改为 proj_in，llm2vae 改为 proj_out，并同步更新 forward 中的引用与 _cast_direct 中的类型转换循环。
3. 用外部 guardrails 包替换自实现：python/sglang/multimodal_gen/runtime/pipelines_core/stages/model_specific_stages/cosmos3_guardrails.py 从 ~400 行重写为 ~60 行，删除 SafetyClassifier、_pixelate_face、_download_checkpoint、_build_text_guardrail 等函数，新增 _init_guardrails（惰性加载 CosmosSafetyChecker 并支持 offload_to_cpu），以及 check_text_safety、check_video_safety 两个外部调用接口；Cosmos3TextGuardrailStage 的 forward 改为调用 check_text_safety；并修复批处理视频检查的 bug。
4. 更新测试：python/sglang/multimodal_gen/test/unit/test_cosmos3.py 中的 test_model_norm_dropped 改为 test_norm_dropped（去掉 model. 前缀），增加 test_audio_proj_in_dropped、test_action_proj_in_dropped；GEN/UND 路径的 Q/K/V 测试键名同步改为新格式，并新增 test_gen_norm_added_q/k。

5. 标记未支持的模态：在参数映射中添加 r"^audio_.*$": "" 和 r"^action_.*$": ""，避免加载时产生警告。

关键文件：

- python/sglang/multimodal_gen/runtime/pipelines_core/stages/model_specific_stages/cosmos3_guardrails.py（模块 安全检查；类别 source；类型 data-contract；符号 SafetyClassifier, init, forward, _pixelate_face）：核心重构：删除全部自实现的安全检查代码，替换为单个外部包调用，文件长度由 413 行缩减至 65 行。
- python/sglang/multimodal_gen/configs/models/dits/cosmos3video.py（模块 模型配置；类别 source；类型 data-contract）：权重映射核心变更，影响所有模型权重加载；删除了 model 前缀，调整了 attention 投影和时间嵌入的键名。
- python/sglang/multimodal_gen/test/unit/test_cosmos3.py（模块 单元测试；类别 test；类型 test-coverage；符号 test_model_norm_dropped, test_norm_dropped, test_audio_proj_in_dropped, test_action_proj_in_dropped）：测试全面更新，验证新映射的正确性，新增对前向兼容的测试用例。
- python/sglang/multimodal_gen/runtime/models/dits/cosmos3video.py（模块 模型运行时；类别 source；类型 data-contract）：潜变量投影层重命名，影响模型结构定义和前向传播。

关键符号：_build_cosmos3_param_names_mapping, _init_guardrails, check_text_safety, check_video_safety, Cosmos3Video.init, Cosmos3Video.forward, Cosmos3TextGuardrailStage.forward

## 评论区精华

gemini-code-assist 指出了两个问题：

- 在 check_video_safety 中，当输入视频维度为 5（批处理）时，只处理了 video[0]，忽略其余批次，导致正确性问题。作者已修复为循环处理所有批次。
- 在 _init_guardrails 中，遍历 runner.models 前应判断 runner 是否为 None 并检查 hasattr，避免潜在的 AttributeError。作者已添加防御性检查。 两个问题均已解决并推送。
- check_video_safety 批处理 bug (correctness): 作者已修复，改为循环处理所有批次并堆叠输出。
- _init_guardrails 防御性检查 (correctness): 作者已添加 if runner is not None and hasattr(runner, "models") 的防御性判断。

## 风险与影响

- 风险：
  - 外部依赖风险：用户必须手动安装 cosmos-guardrail==0.3.1，若遗漏则运行时抛出 ImportError；该包可能随着 API 更新而破坏兼容性。
  - 权重映射兼容性：旧版本 checkpoint 因前缀和键名变化无法直接加载，需要重新转换权重或使用迁移脚本。

    - 批处理安全性修复有效性：虽然作者已修改批处理逻辑，但测试未覆盖多视频场景，可能仍有边界条件未覆盖。
    - 模态预留：audio/action 权重被显式跳过，若后续需要支持，需在映射中添加对应条目并更新模型结构。
- 影响：
    - 用户：用户需要额外执行 pip install cosmos-guardrail==0.3.1；如果使用旧 checkpoint 或自定义加载流程，需调整权重键名。
    - 系统：安全检查不再由 SGLang 内部维护，而是委托给外部包，减轻了维护负担，但也引入了对第三方包的依赖。
    - 团队：Cosmos3 相关代码量显著减少（-495 行），架构更清晰；但需要关注外部包的更新和兼容性。
    - 风险标记：外部依赖 cosmos_guardrail，权重映射破坏旧 checkpoint 兼容性，批处理 bug 修复需验证，音频 / 动作模态未支持

## 关联脉络

- PR #24994 Initial support for new world model: 本 PR 是对 #24994 初始支持的后续更新，适配最终 checkpoint 格式并重构 guardrails。