

PR #48211 完整报告

vllm-project/vllm

[Model][CI/Build] Cosmos3: enable registry tests and register Cosmos3-Super

合并时间: 2026-07-10 16:35

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/48211>

执行摘要

- 一句话: 启用 Cosmos3 注册测试并注册 Super 变体
- 推荐动作: 建议合入。该 PR 清除了未及时清理的暂存标记, 完善了 Cosmos3 测试覆盖和文档, 逻辑清晰且风险极低。

功能与动机

PR #43356 合并时因 checkpoints 未发布而设置了 `is_available_online=False`, 但后续 `nvidia/Cosmos3-Nano` 已上线, 该标记导致注册驱动测试 (初始化、多模态处理、张量 schema) 仍跳过该架构。

实现拆解

1. 移除 `is_available_online=False`: 在 `tests/models/registry.py` 中删除该标记, 使 Cosmos3 能参与 CI 注册测试。
2. 注册 `nvidia/Cosmos3-Super`: 在 `extras` 中增加 `super` 变体, 其模型结构与 Nano 一致但更深 (64 层 vs 36 层)。
3. 显式跳过 VLM 生成测试: 由于 Transformers 无 `cosmos3_omni` 映射, HF 参考 runner 无法加载 checkpoint, 因此在 `tests/models/multimodal/generation/test_common.py` 中添加 `pytest.mark.skip` 标记。
4. 更新文档: 在 `docs/models/supported_models.md` 的 Cosmos3 行中加入 `nvidia/Cosmos3-Super`。

关键文件:

- `tests/models/registry.py` (模块 注册表; 类别 test; 类型 test-coverage) : 移除了 `is_available_online=False` 并注册了 `Cosmos3-Super` 作为 `extras`, 是启用 CI 测试的关键变更。
- `tests/models/multimodal/generation/test_common.py` (模块 测试配置; 类别 test; 类型 test-coverage) : 显式跳过 VLM 生成测试, 避免因 Transformers 不支持 `cosmos3_omni` 而导致 CI 失败。
- `docs/models/supported_models.md` (模块 文档; 类别 docs; 类型 documentation) : 文档更新, 列出 `Cosmos3-Super` 变体。

关键符号: 未识别

关键源码片段

tests/models/registry.py

移除了 `is_available_online=False` 并注册了 Cosmos3-Super 作为 extras，是启用 CI 测试的关键变更。

```
# tests/models/registry.py 中 Cosmos3 注册条目变更
# 变更前: is_available_online=False 导致所有注册驱动测试被跳过
"Cosmos3ForConditionalGeneration": _HfExamplesInfo(
    "nvidia/Cosmos3-Nano",
    max_model_len=4096,
    min_transformers_version="4.57",
    is_available_online=False, # 旧: 标记移除前
),
# 变更后: 移除标记并注册 Super 变体
"Cosmos3ForConditionalGeneration": _HfExamplesInfo(
    "nvidia/Cosmos3-Nano",
    extras={"super": "nvidia/Cosmos3-Super"}, # 新: 注册更深版本的 Super
    max_model_len=4096,
    min_transformers_version="4.57",
),
```

tests/models/multimodal/generation/test_common.py

显式跳过 VLM 生成测试，避免因 Transformers 不支持 `cosmos3_omni` 而导致 CI 失败。

```
# tests/models/multimodal/generation/test_common.py 中 cosmos3 配置变更
"cosmos3": VLMTestInfo(
    models=["nvidia/Cosmos3-Nano"],
    # ... 其他配置保持不变 ...
    marks=[ # 新增: 显式跳过生成测试
        pytest.mark.skip(
            reason="Transformers has no cosmos3_omni mapping, so the HF "
            "reference runner cannot load the checkpoint."
        )
    ],
),
```

评论区精华

无实质讨论。PR 由 [DarkLight1337](#) 批准，无 review 评论。

- 暂无高价值评论线程

风险与影响

- 风险：低风险。变更仅涉及测试配置和文档，无核心逻辑改动。移除 `is_available_online=False` 后，CI 会真实加载 Cosmos3 checkpoint，若实际权重有差异可能导致测试失败，但 PR 作者已手动验证真实模型输出正确。Super 变体仅参与 processor 测试，不会初始化完整模型。

- 影响：影响范围小，仅限于测试和文档。Cosmos3 注册测试（初始化、多模态处理）将不再是 skip 状态，而是实际执行。开发者可在 CI 中获得 Cosmos3 架构的测试覆盖；文档列出 Super 变体有助于用户知晓支持情况。
- 风险标记：缺少测试覆盖

关联脉络

- PR #43356 [Model] Add Cosmos3 support: 为本 PR 的前置 PR，引入了 Cosmos3 支持并设置了 is_available_online=False。本 PR 在该 PR 基础上启用测试并注册 Super 变体。