

PR #52064 完整报告

vllm-project/vllm

[CI] Mirror external test assets in vLLM S3

合并时间: 2026-08-13 10:08

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/52064>

执行摘要

- 一句话: 将外部测试资源镜像到 vLLM S3, 消除 CI 对第三方主机依赖
- 推荐动作: 值得快速浏览: 本 PR 展示了如何系统性消除 CI 对第三方主机的依赖, 其中 "统一 VLLM_S3_BUCKET_URL 常量 + 动态推导 SSRF 白名单" 的做法值得借鉴。若后续出现新的外部资源依赖, 可按同样模式迁移。

功能与动机

PR body 明确指出: "Buildkite build 83608 had multiple test failures caused by direct dependencies on third-party hosts, including RemoteDisconnected while fetching OpenCV and Bogotobogo video fixtures"。测试直接依赖第三方站点导致 CI 不稳定, 因此将资源镜像到 vLLM 自有 S3 桶, 并验证镜像对象与源文件逐字节一致。

实现拆解

1. 统一资源前缀: 在 5 个测试文件中引入 `from vllm.assets.base import VLLM_S3_BUCKET_URL`, 将原本硬编码的第三方 URL / S3 URL 统一改为由该常量拼接, 便于后续统一迁移。
2. 替换 JinaVL 重排序图片 (`tests/models/multimodal/pooling/test_jinavl_reranker.py`): 新增 `HANDELSBLATT_IMAGE_URL`、`PAPER_IMAGE_URL` 常量, 替换 `TEXT_IMAGE_TEST_DATA`、`IMAGE_TEXT_TEST_DATA`、`IMAGE_IMAGE_TEST_DATA`、`TEXT_MIXED_DOCS_TEST_DATA` 中 4 处指向 jina-ai GitHub raw 的图片地址, 改动量最大 (+14/-28)。
3. 替换视频资源 (`tests/multimodal/media/test_connector.py`、`tests/entrypoints/multimodal/openai/chat_completion/test_video.py`、`tests/entrypoints/pooling/classify/test_online_vision.py`): 将 Bogotobogo 与 OpenCV GitHub 的视频 URL 改为 `{VLLM_S3_BUCKET_URL}/multimodal_asset/` 下的 `slow_traffic_small.mp4`、`vtest.avi`、`Megamind.avi`; 同时 `test_online_vision.py` 中原本硬编码的 S3 图片地址也改为常量拼接。
4. 调整 SSRF 域白名单 (`tests/multimodal/media/test_connector.py`): 由于资源域从 `bogotobogo.com` / `github.com` 变为 `vllm-public-assets`, `allowed_media_domains` 改为 `VLLM_S3_BUCKET_URL.removeprefix("https://")` 动态推导, 避免白名单与资源地址脱节。

5. 替换 GSM8K 数据集 (tests/evals/gsm8k/gsm8k_eval.py) : load_gsm8k_data 中 train/test 的 GitHub raw URL 改为 {VLLM_S3_BUCKET_URL}/ci-datasets/gsm8k/, 所有基于该脚本的 eval CI 都会受益。
6. 验证配套: PR body 说明 7 个镜像资源均从公共 S3 获取并与源 SHA256 校验一致, OpenCV 成功解码全部视频 (Megamind.avi 270 帧、slow_traffic_small.mp4 914 帧、vtest.avi 795 帧) , GSM8K 数据逐行 JSON 解析通过。

关键文件:

- tests/models/multimodal/pooling/test_jinavl_reranker.py (模块 重排序; 类别 test; 类型 test-coverage) : 改动量最大的文件, 将 jina-ai GitHub raw 的 4 处图片 URL 替换为 S3 镜像并抽取为常量, 是本次镜像工作的代表。
- tests/multimodal/media/test_connector.py (模块 媒体连接; 类别 test; 类型 test-coverage) : 除视频 URL 镜像外, 还将 allowed_media_domains 从硬编码外部域改为由 VLLM_S3_BUCKET_URL 动态推导, 保证 SSRF 域白名单与资源地址一致。
- tests/evals/gsm8k/gsm8k_eval.py (模块 评测脚本; 类别 test; 类型 test-coverage) : GSM8K train/test 数据源从 GitHub raw 切到 S3 ci-datasets 目录, 影响所有基于该脚本的 eval CI。
- tests/entrypoints/multimodal/openai/chat_completion/test_video.py (模块 视频测试; 类别 test; 类型 test-coverage) : OpenAI 兼容入口的视频测试改用 S3 镜像, 覆盖三个视频资源。
- tests/entrypoints/pooling/classify/test_online_vision.py (模块 在线视觉; 类别 test; 类型 test-coverage) : 图像与视频 URL 统一改用常量, 移除硬编码 S3 地址与第三方视频地址。

关键符号: load_gsm8k_data, test_allowed_media_domains, test_fetch_image_base64

关键源码片段

tests/models/multimodal/pooling/test_jinavl_reranker.py

改动量最大的文件, 将 jina-ai GitHub raw 的 4 处图片 URL 替换为 S3 镜像并抽取为常量, 是本次镜像工作的代表。

```
from vllm.assets.base import VLLM_S3_BUCKET_URL

# 原来直接引用 jina-ai 的 GitHub raw 图片, 受第三方主机稳定性影响;
# 镜像后统一走 S3, 常量集中管理, 测试数据里只引用常量。
HANDELSBLATT_IMAGE_URL = (
    f"{VLLM_S3_BUCKET_URL}/multimodal_asset/jinavl-handelsblatt-preview.png"
)
PAPER_IMAGE_URL = f"{VLLM_S3_BUCKET_URL}/multimodal_asset/jinavl-paper-11.png"

IMAGE_IMAGE_TEST_DATA = {
    "query": [{"image": PAPER_IMAGE_URL}],
    "documents": [
        {"image": HANDELSBLATT_IMAGE_URL},
        {"image": PAPER_IMAGE_URL},
    ],
}
```

```

    ],
}

```

tests/multimodal/media/test_connector.py

除视频 URL 镜像外，还将 `allowed_media_domains` 从硬编码外部域改为由 `VLLM_S3_BUCKET_URL` 动态推导，保证 SSRF 域白名单与资源地址一致。

```

from vllm.assets.base import VLLM_S3_BUCKET_URL

# 视频资源从第三方站点迁移到 vLLM 公共 S3 桶，
# URL 前缀统一由常量推导，避免硬编码第三方域名再次失效。
TEST_VIDEO_URLS = [
    f"{VLLM_S3_BUCKET_URL}/multimodal_asset/slow_traffic_small.mp4",
    f"{VLLM_S3_BUCKET_URL}/multimodal_asset/vtest.avi",
]

@pytest.mark.asyncio
@pytest.mark.parametrize("video_url", TEST_VIDEO_URLS)
@pytest.mark.parametrize("num_frames", [-1, 32, 1800])
async def test_allowed_media_domains(video_url: str, num_frames: int):
    connector = MediaConnector(
        media_io_kwargs={"video": {"num_frames": num_frames}},
        # 资源迁移到 S3 后，白名单也改为从常量动态推导，
        # 避免域名变更时域白名单与资源地址脱节。
        allowed_media_domains=[VLLM_S3_BUCKET_URL.removeprefix("https://")],
    )
    # 同步与异步取流结果必须一致
    video_sync, metadata_sync = connector.fetch_video(video_url)
    video_async, metadata_async = await connector.fetch_video_async(video_url)
    assert np.array_equal(video_sync, video_async)
    assert metadata_sync == metadata_async

```

tests/evals/gsm8k/gsm8k_eval.py

GSM8K train/test 数据源从 GitHub raw 切到 S3 ci-datasets 目录，影响所有基于该脚本的 eval CI。

```

def load_gsm8k_data() -> tuple[list[dict], list[dict]]:
    """Load GSM8K train and test data"""
    # GSM8K 数据镜像到公共 S3，避免直接依赖 GitHub raw 的可用性；
    # 前缀从 VLLM_S3_BUCKET_URL 常量推导，便于统一更换存储位置。
    train_url = f"{VLLM_S3_BUCKET_URL}/ci-datasets/gsm8k/train.jsonl"
    test_url = f"{VLLM_S3_BUCKET_URL}/ci-datasets/gsm8k/test.jsonl"

    train_file = download_and_cache_file(train_url)
    test_file = download_and_cache_file(test_url)

    train_data = list(read_jsonl(train_file))
    test_data = list(read_jsonl(test_file))

```

```
return train_data, test_data
```

评论区精华

该 PR 来自 fork, claude[bot] 提示自动 review 被禁用; 维护者 AndreasKaratzas 与 njhill 均直接 approve ("LGTM"、"Thanks @khluu !")。PR body 中作者主动说明已检索过相关 PR, 确认无重复工作, 且 #52043 属于独立的混合依赖构建问题, 刻意排除在本 PR 之外。全程无实质技术争议。

- Fork 提交的自动 review 限制 (other): 无技术争议, 维护者均认可变更。

风险与影响

- 风险:
 1. 公共 S3 桶可用性依赖: 镜像资源位于 vllm-public-assets (us-west-2), 若该桶配置、权限或对象内容出现问题, 相关测试会失败; 缺少对 S3 资源存在性的 CI 预检。
 2. SSRF 域白名单收窄: test_connector.py 的 allowed_media_domains 从两个外部域改为仅允许 S3 域, 若未来测试需要引用其他域, 需要重新扩展白名单逻辑。
 3. 区域网络延迟: us-west-2 之外运行的 CI 拉取视频可能比原先的 CDN 慢, 但通常比第三方主机更稳定; ROCm 测试已保留超时放大设置。
 4. 无生产代码变更: 所有改动限于 tests/ 目录, 不会引入运行时回归风险。- 影响: 对用户无影响 (纯测试基础设施变更); 对 CI 稳定性有明显改善, 消除了视频、图片、GSM8K 数据集对第三方主机的网络依赖; 对团队而言, 后续新增测试资源需要遵循 "镜像到 S3 + 使用 VLLM_S3_BUCKET_URL 常量" 的约定, 并维护镜像与源文件的校验机制。- 风险标记: 依赖公共 S3 桶可用性, 域白名单收窄, 测试基础设施变更

关联脉络

- PR #52043 [CI] Force source builds for hybrid dependencies: PR body 明确提及这是另一个独立的混合依赖构建失败, 刻意排除在本 PR 之外; 两者同属 CI 稳定性改进。
- PR #52028 [Bugfix] Pin DeepEP by its full commit hash: 同属消除外部依赖不稳定性的 CI 基础设施修复, 与 S3 镜像目标一致。