PR #52570 完整报告

vllm-project/vllm

[CI/Build] Reduce more duplicate runner startup in tests

合并时间：2026-08-17 13:19

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/52570

执行摘要

- 一句话：合并多项池化测试启动，减少约 53 次 runner 启动
- 推荐动作：值得测试相关的工程师精读，特别是 test_colpali.py 的 test_colpali_default_runner 聚合模式、test_colqwen3_5.py 的 module fixture 模式，以及 test_awq.py 把参数化改为循环的手法。可作为 vLLM 测试套件中「减少 runner 启动」系列重构的参考实现。若需进一步优化，可考虑为共享 fixture 增加失败隔离（如按 fixture 分组标记）并保留异常消息断言。

功能与动机

PR body 明确说明这是 #52417 的 follow-up，目标是在更多 generation/pooling 测试中减少 runner 启动。表格显示 AWQ size factors、Whisper beam widths、ColBERT、Truncation control、CLIP、SigLIP、ColModernVBERT、ColPali、ColQwen3.5、Llama Nemotron 等测试组共节省 53 次启动。每减少一次 vLLM runner 启动，就少一次模型权重加载和引擎初始化，直接降低 CI 耗时与 GPU 资源占用。

实现拆解

本 PR 全部改动位于 tests/ 目录，核心手法是「把 pytest 参数化拆分的多个 runner 生命周期合并为单次启动内的多 case 执行」，具体分步拆解如下：

1. 共享 runner 生命周期（module 级 fixture）：test_colqwen3_5.py、test_colmodernvbert.py、test_colbert.py、test_truncation_control.py 均新增 scope="module" 的 fixture（如 colqwen3_5_model、colmodernvbert_model、colbert_model、vllm_model），把原先每个测试函数内 with vllm_runner(...) 的启动逻辑上移，多个测试函数改为直接使用同一模型实例。test_colbert.py 中 test_colbert_late_interaction_1_to_1/1_to_N/N_to_N、test_colbert_relevance_ordering、test_colbert_embed_not_supported 等 6 个测试全部改从 colbert_model fixture 取值，一次启动覆盖全部用例。需注意 test_colbert_embed_not_supported 原本把 pytest.raises 与 runner 一起作为 with 上下文，现在改为在共享实例上断言 ValueError，验证异常后 engine 仍可用。
2. 拆分测试合并为单 case 列表：test_llama_nemotron_vl.py、test_siglip.py、test_clip.py 将原本 test_models_text + test_models_image 两个独立测试合并为 test_models，_run_test 改为接收 input_cases 列表，在一个 vLLM runner 和一个 HF runner 生命周期内循环跑 text/image 两组 case，最后逐组 check_embeddings_close。

test_llama_nemotron_vl.py 的 reranker 部分同样把 _run_hf_reranker、_run_vllm_reranker 改成接收 Sequence[RerankerCase]，并新增 RerankerDocument/RerankerCase 类型别名，text-only 与 image-doc 两类 rerank case 在一次启动内完成。

3. ColPali 聚合测试 + V2 runner 专项保留：test_colpali.py 把 _run_token_embed_test、_run_late_interaction_test、_run_relevance_test、_run_multimodal_mixed_docs_test、_run_multimodal_image_query_text_docs_test 的 runner 启动全部抽取，新增 test_colpali_default_runner 在单次启动内串行执行全部 5 项子测试（vLLM runner 启动 24 → 4）。同时保留 test_colpali_v2_multimodal_text_query_image_docs，专门用 VLLM_USE_V2_MODEL_RUNNER=1 环境变量 + FLASH_ATTN 后端 + 关闭 flashinfer autotune 覆盖 MRV2 路径，因为该配置与默认 runner 不同，无法共享实例。
4. 参数化转循环（AWQ）：test_awq.py 把 size_factors 从 pytest parametrize（每个 factor 组一个 runner 生命周期）改为函数内常量 IMAGE_SIZE_FACTOR_GROUPS 三重循环（单尺度、批处理、多尺度），一次启动内生成 6 组输入并对比 source/quant 模型输出，vLLM runner 启动 6 → 2。
5. Whisper beam 宽度内聚：test_whisper.py 定义 BEAM_WIDTHS = (1, 2)，移除 beam_width 参数化装饰器，改为在单次 HF/vLLM runner 生命周期内分别对两种 beam width 调用 generate_beam_search，两次启动合并为一次；断言同时也更严格（len(hf_output_ids) == len(vllm_output_ids) == beam_width）。
6. 配套调整：test_truncation_control.py 在改用共享 fixture 后，test_bigger_truncation_size 移除了原先对异常消息的逐字断言（只保留 pytest.raises(VLLMValidationError)），属于为共享实例让路的断言弱化；CLIP/SigLIP 的 test_models_text_image_no_crash 逻辑（同时传 text+image 应抛 ValueError，且之后请求仍可用）被内联进 _run_test，随单次启动一并验证。

所有改动均为测试实现层面的重构，不涉及 vllm/ 下源码、配置或部署文件。

关键文件：

- tests/models/multimodal/pooling/test_llama_nemotron_vl.py（模块 多模态池化；类别 test；类型 test-coverage；符号 test_models_text, test_models_image, test_models, _pil_to_data_uri）：改动最大的文件（+150/-171），把 text/image embedding 两组测试合并为单次 runner 生命周期，并新增 RerankerDocument/RerankerCase 类型别名让 reranker 多 case 单次启动
- tests/models/multimodal/pooling/test_colpali.py（模块 多模态池化；类别 test；类型 test-coverage；符号 test_colpali_token_embed, test_colpali_late_interaction_scoring, test_colpali_relevance_ordering, test_colpali_default_runner）：vLLM 启动 24 → 4 的核心：新增 test_colpali_default_runner 在单次启动内串行跑 5 项子测试，同时保留 V2 runner 专项测试覆盖 MRV2 路径
- tests/models/quantization/test_awq.py（模块 量化测试；类别 test；类型 test-coverage）：把 size_factors 从 pytest 参数化改为函数内 IMAGE_SIZE_FACTOR_GROUPS 循环，vLLM runner 启动 6 → 2，展示参数化转循环的典型手法
- tests/models/multimodal/pooling/test_siglip.py（模块 多模态池化；类别 test；类型 test-coverage；符号 test_models_text, test_models, test_models_image,

test_models_text_image_no_crash）：test_models_text + test_models_image +test_models_text_image_no_crash 合并为单个 test_models，vLLM 9 → 3、HF 6 → 3

- tests/models/multimodal/pooling/test_clip.py（模块 多模态池化；类别 test；类型 test-coverage；符号 test_models_text, test_models_image, test_models, test_models_text_image_no_crash）：与 SigLIP 相同模式：text/image 测试合并为 test_models，no-crash 校验内联，vLLM 3 → 1、HF 2 → 1
- tests/models/multimodal/pooling/test_colqwen3_5.py（模块 多模态池化；类别 test；类型 test-coverage；符号 _run_token_embed_test, colqwen3_5_model, test_colqwen3_5_token_embed, test_colqwen3_5_late_interaction_scoring）：引入 module 级 colqwen3_5_model fixture，三个测试共享一次启动，vLLM 3 → 1
- tests/models/language/pooling/test_colbert.py（模块 语言池化；类别 test；类型 test-coverage；符号 test_colbert_token_embed, colbert_model）：ColBERT 6 个测试共享 colbert_model fixture，vLLM 启动 24 → 4 的主要贡献者之一
- tests/models/language/pooling/test_truncation_control.py（模块 语言池化；类别 test；类型 test-coverage；符号 vllm_model, test_max_truncation_size, test_bigger_truncation_size）：共享 vllm_model fixture 合并 4 个截断测试启动，但删除了异常消息的逐字断言，是断言强度弱化的案例
- tests/models/multimodal/generation/test_whisper.py（模块 多模态生成；类别 test；类型 test-coverage）：beam_width 参数化（2 次启动）改为 BEAM_WIDTHS 循环（1 次启动），并收紧 beam 数断言
- tests/models/multimodal/pooling/test_colmodernvbert.py（模块 多模态池化；类别 test；类型 test-coverage；符号 test_colmodernvbert_text_token_embed, colmodernvbert_model, test_colmodernvbert_text_relevance_ordering, test_colmodernvbert_text_late_interaction）：module 级 colmodernvbert_model fixture 合并 4 个测试启动，vLLM 4 → 1

关键符号：_run_test, _run_hf_reranker, _run_vllm_reranker, _run_token_embed_test, _run_late_interaction_test, _run_relevance_test, _run_multimodal_text_query_image_docs_test, _run_multimodal_mixed_docs_test, _run_multimodal_image_query_text_docs_test, run_awq_test, test_beam_search_encoder_decoder

关键源码片段

tests/models/multimodal/pooling/test_llama_nemotron_vl.py

改动最大的文件（+150/-171），把 text/image embedding 两组测试合并为单次 runner 生命周期，并新增 RerankerDocument/RerankerCase 类型别名让 reranker 多 case 单次启动

```
# tests/models/multimodal/pooling/test_llama_nemotron_vl.py
# 核心重构：_run_test 不再只跑一组输入，而是接收 input_cases 列表，
# 在一个 vLLM runner 与一个 HF runner 生命周期内完成全部对比，
# 从而把原来 test_models_text / test_models_image 两次启动合并为一次。

def _run_test(
```

```
    hf_runner: type[HfRunner],
    vllm_runner: type[VllmRunner],
    input_cases: list[tuple[list[str], PromptImageInput]],
    model: str,
    *,
    dtype: str,
) -> None:
    """Compare HF and vLLM embeddings for all input cases.

    NOTE: Run vLLM first to avoid CUDA initialization issues with multiprocessing.
    """
    # 单次 vLLM 启动，循环跑所有 case（text-only / image）
    with vllm_runner(
        model,
        runner="pooling",
        dtype=dtype,
        max_model_len=2048,
        enforce_eager=True,
        trust_remote_code=True,
        **ROCM_ENGINE_KWARGS,
    ) as vllm_model:
        vllm_outputs_per_case = [
            vllm_model.embed(input_texts, images=input_images)
            for input_texts, input_images in input_cases
        ]

    # 单次 HF 启动，同样循环跑所有 case
    with hf_runner(model, dtype=dtype, auto_cls=AutoModel) as hf_model:
        hf_outputs_per_case = []
        for input_texts, input_images in input_cases:
            hf_outputs = []
            for text, image in zip(input_texts, input_images):
                with torch.inference_mode():
                    # 按 query/passage 前缀区分调用 encode_queries 或 encode_documents
                    if text.startswith(QUERY_PREFIX):
                        query_text = text[len(QUERY_PREFIX):]
                        embedding = hf_model.model.encode_queries([query_text])
                    elif text.startswith(PASSAGE_PREFIX):
                        passage_text = text[len(PASSAGE_PREFIX):]
                        if image is not None:
                            embedding = hf_model.model.encode_documents(
                                images=[image], texts=[passage_text])
                        else:
                            embedding = hf_model.model.encode_documents(
                                texts=[passage_text])
                    else:
                        raise ValueError(
                            f"Text must start with {QUERY_PREFIX!r} "
                            f"or {PASSAGE_PREFIX!r}")
```

```
                hf_outputs.append(embedding[0].tolist())
            hf_outputs_per_case.append(hf_outputs)

    # 逐 case 对比，保持原有校验强度不变
    for hf_outputs, vllm_outputs in zip(hf_outputs_per_case, vllm_outputs_per_case):
        check_embeddings_close(
            embeddings_0_lst=hf_outputs,
            embeddings_1_lst=vllm_outputs,
            name_0="hf",
            name_1="vllm",
        )
```

tests/models/multimodal/pooling/test_colpali.py

vLLM 启动 24 → 4 的核心：新增 test_colpali_default_runner 在单次启动内串行跑 5 项子测试，同时保留 V2 runner 专项测试覆盖 MRV2 路径

```
# tests/models/multimodal/pooling/test_colpali.py
# 聚合测试：原来 24 次 vLLM runner 启动（多模型 × 多测试）压缩为
# 默认 runner 一次 + V2 runner 一次，所有校验在 helper 内保持等价。

@pytest.mark.parametrize("model", MODELS)
@pytest.mark.parametrize("dtype", [DTYPE])
def test_colpali_default_runner(
    vllm_runner,
    model: str,
    dtype: str,
) -> None:
    # 单次启动，依次执行 5 项子测试，复用同一 engine 实例
    with vllm_runner(
        model,
        runner="pooling",
        dtype=dtype,
        max_model_len=4096,
        enforce_eager=True,
        gpu_memory_utilization=GPU_MEMORY_UTILIZATION,
    ) as vllm_model:
        _run_token_embed_test(vllm_model, model)
        _run_late_interaction_test(vllm_model)
        _run_relevance_test(vllm_model)
        _run_multimodal_mixed_docs_test(vllm_model)
        _run_multimodal_image_query_text_docs_test(vllm_model)


@pytest.mark.parametrize("model", MODELS)
@pytest.mark.parametrize("dtype", [DTYPE])
def test_colpali_v2_multimodal_text_query_image_docs(
    vllm_runner,
    monkeypatch: pytest.MonkeyPatch,
    model: str,
```

```
    dtype: str,
) -> None:
    # V2 runner 使用不同 backend/autotune 配置，无法与默认 runner 共享实例，
    # 因此单独保留一次启动，专门验证 MRV2 下的多模态打分路径
    monkeypatch.setenv("VLLM_USE_V2_MODEL_RUNNER", "1")
    with vllm_runner(
        model,
        runner="pooling",
        dtype=dtype,
        max_model_len=4096,
        enforce_eager=True,
        gpu_memory_utilization=GPU_MEMORY_UTILIZATION,
        attention_backend="FLASH_ATTN",
        kernel_config={"enable_flashinfer_autotune": False},
    ) as vllm_model:
        assert vllm_model.llm.llm_engine.vllm_config.use_v2_model_runner
        _run_multimodal_text_query_image_docs_test(vllm_model)
```

评论区精华

该 PR 的 review 讨论很少：仅有一条 Claude bot 的提示（fork 的 PR 默认禁用自动 review，维护者可手动触发），以及合并者 DarkLight1337 的 approved 评论 "Thanks cc @noooop"（@noooop 可能是 ColPali/ 池化相关维护者）。此外作者触发了一次 Codex Review，结论是 "Didn't find any major issues"。没有 reviewer 对合并策略、fixture 作用域或断言弱化提出质疑。值得注意的是测试本身隐含一个设计权衡：共享 runner 会让单个测试失败影响同 fixture 的其他用例，且 test_colbert_embed_not_supported 在共享实例上验证异常后，后续测试若复用该实例可能受异常后状态影响——但当前文件内该测试是最后执行的，风险可控。

- fork PR 的自动 review 策略 (other)：由维护者 DarkLight1337 直接 approve 并合并，无人工 review 争议。
- 合并确认与维护者知会 (question)：合并者认可变更，无进一步修改要求。

风险与影响

- 风险：主要风险集中在测试语义变化而非源码回归：
  1. 断言等级变化：test_colbert_embed_not_supported 从「异常 +runner 上下文」改为共享实例上断言异常；test_truncation_control.py 删除了 test_bigger_truncation_size 的异常消息逐字比较，错误信息回归可能漏检。
  2. 耦合度上升：module 级 fixture 使多个测试共享一个 engine 实例，单个测试内部产生异常或污染状态会波及其他用例；test_colpali_default_runner 串行执行 5 项子测试，失败定位粒度变粗（虽然断言消息仍能区分）。
  3. 行为保持性：test_colpali_v2_multimodal_text_query_image_docs 使用不同 backend 配置，无法与默认测试共享实例，属有意保留；若未来默认 runner 切换为 MRV2，该覆盖可能会退化。

4. Whisper 断言收紧：新增 len(hf_output_ids) == len(vllm_output_ids) == beam_width，如果 HF 端在某些输入下返回的 beam 数少于请求宽度，测试会开始失败（这属于更严格，通常可接受）。
5. 所有风险都限于 tests/ 目录，不影响 vLLM 运行时行为。 - 影响：影响范围限定在 CI 测试基础设施：vLLM runner 启动 64 → 18、HF runner 14 → 7，合计约 53 次启动，直接降低 Buildkite CI 耗时和 GPU/ 内存资源占用，对多模态池化测试（ColPali、ColQwen、ColModernVBERT、Llama Nemotron VL、CLIP、SigLIP）与 ColBERT/Whisper/AWQ 测试的每次合并请求 CI 都有收益。对用户无功能影响；对团队意味着后续新增池化测试时有更高效的「多 case 单 runner」范式可参考，但共享 fixture 也会增加用例间耦合。 - 风险标记：纯测试改动，共享 fixture 耦合，断言弱化，CI 耗时优化

关联脉络

- PR #52417 [CI/Build] Avoid duplicate runner startup for multimodal test: 本 PR 的 direct follow-up，同一作者同一主题，前序已合并多模态 generation 测试的 runner 启动，本 PR 扩展到 pooling/generation 更多测试组
- PR #52050 [Bugfix] Temporarily disable FA4 head-dim 256: 涉及 ColPali 在 MRV2 下的崩溃修复，本 PR 中 test_colpali_v2_multimodal_text_query_image_docs 保留的 V2 runner 配置（FLASH_ATTN）与该修复相关
- PR #52425 [ModelRunner v2] Support Transformers pooling model: MRV2 池化模型支持相关，本 PR 合并 ColPali/ColQwen 测试启动时保留 V2 runner 专项目覆盖
- PR #52246 [Bugfix][Anthropic] Return 4xx for client-caused errors in /v1/messages: 同一 author 的近期前端测试相关工作，反映 contributors 在 CI/ 测试效率与 frontend 错误处理上的并行投入