PR #39607 完整报告

vllm-project/vllm

[Doc] Add Gemma 4 to supported models list

合并时间：2026-04-17 13:42

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/39607

执行摘要

本 PR 是一个文档更新，将已实现的 Gemma 4 模型（Gemma4ForCausalLM 和 Gemma4ForConditionalGeneration）添加到 vLLM 的支持模型列表文档中，并澄清了音频和视频模态的支持细节，旨在提高文档准确性和用户体验，无代码变更，风险极低。

功能与动机

为什么做：Gemma 4 模型已在 vLLM 代码中实现（如 gemma4.py、gemma4_mm.py），但未在官方文档 supported_models.md 中列出，导致用户可能无法发现或误用。PR body 明确说明："Gemma4ForCausalLM and Gemma4ForConditionalGeneration are already implemented and registered in vLLM, but were missing from the documentation table"，因此需要更新文档以反映实际支持。

实现拆解

变更仅涉及一个文件 docs/models/supported_models.md，按以下步骤拆解：

1. 更新文本模型表格：在"Text-only models"部分添加 Gemma4ForCausalLM 行，包括模型名、描述、示例 Hugging Face ID（如 google/gemma-4-E2B-it）、LoRA 和管道并行（PP）支持状态，格式与现有 Gemma 3 条目一致。
2. 更新多模态模型表格：在"Multimodal models"部分添加 Gemma4ForConditionalGeneration 行，指定模态为 T + I⁺ + V + A*（文本、多图像、视频、音频），其中 * 标记表示音频仅特定变体支持，并标注 PP 支持。
3. 添加新注释和通用标记：在文档脚注中新增 [*] Only specific variants of the model support this modality 作为可复用通用标记；同时添加 note 块，详细说明音频仅限 gemma-4-E2B 和 gemma-4-E4B 变体，视频非原生输入但 vLLM 实现内部处理，用户可直接发送视频消息。

关键源码片段（整理自文档更新）：

| `Gemma4ForCausalLM` | Gemma 4 | `google/gemma-4-E2B-it`, etc. | ✅✅ | ✅✅ |
| `Gemma4ForConditionalGeneration` | Gemma 4 | T + I⁺ + V + A[*] | `google/gemma-4-E2B-it`, etc. | | ✅✅ |

[*] Only specific variants of the model support this modality (see notes below).

!!! note

For `Gemma4ForConditionalGeneration`:
- audio input is only supported by the `gemma-4-E2B` and `gemma-4-E4B` variants.
- The model does not ingest videos directly. However, vLLM's Gemma 4 implementation supports video inputs by handling video processing internally. Users can send videos directly in the message structure to vLLM.

评论区精华

review 讨论中最有价值的交锋围绕模态支持准确性：

- DarkLight1337初始提问："I think we also have multi-video and multi-audio support"，引发对代码 get_supported_mm_limits 的检查。
- lucianommartins澄清核心细节："all models support text + image; only e2b and e4b support audio input; no model support video input"，并解释视频通过 vLLM 内部 ingestor 处理。
- ywang96基于代码引用确认："Based on gemma4_mm.py lines 207-212, we can support multiple audio"，推动文档更新。
- 最终共识：文档移除视频原生支持标记，添加变体特定音频注释，确保用户角度的准确性。

风险与影响

风险：作为纯文档更新，无技术回归、性能或安全风险；主要风险是文档不准确可能导致用户误解Gemma 4模态支持，但已通过review讨论和代码验证（如引用gemma4_mm.py）降低。
影响：正面影响文档用户，提高 Gemma 4 模型可发现性和使用指导；不影响系统运行；新增通用标记 * 为未来模型文档提供可复用模式，提升团队维护效率。

关联脉络

与历史 PR #39234（"[Models][Gemma4] Prevent GPU/CPU sync in embed_input_ids"）相关，后者修复 Gemma 4 多模态模型的 GPU/CPU 同步问题，本 PR 则补充了该模型的官方文档支持，共同构成 Gemma 4 在 vLLM 中从实现到文档的完整支持链条。近期 PR 趋势显示 vLLM 持续扩展模型支持（如 Gemma 系列、量化整合），本 PR 是这一演进中的文档配套更新。