PR #22729 完整报告

sgl-project/sglang

[Bugfix] Fix Hunyuan3D-2 DiT checkpoint param mapping

合并时间：2026-05-20 10:37

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22729

执行摘要

- 一句话：修复 Hunyuan3D-2 DiT 模型 checkpoint 参数名映射
- 推荐动作：该 PR 值得精读，尤其是 param_names_mapping 的设计模式（正则替换 + merge_info）可复用于其他 DiT 或 Flux 系列模型。建议后续增加参数映射的单元测试，覆盖主要 checkpoint 变体。

功能与动机

Loading the official Hunyuan3D-2 DiT checkpoint with the existing mapping fails with an unsupported "new parameter" error for fused text attention weights, for example: ValueError: New parameter 'double_blocks.10.txt_attn.qkv.weight' is not supported. Similar issues show up for single-stream blocks when the checkpoint uses separate proj_out / attn.to_out projections instead of the packed linear2 layout used by Hunyuan3D2DiT.

实现拆解

1. 添加模型前缀去除规则：在映射字典开头添加 r"^model\.(.*)$" → r"\1"，以兼容官方 safetensors 中可能存在的 model. 前缀。
2. MLP 线性层重命名：将 double_blocks 中索引为 0 和 2 的 MLP 层（分别对应 fc_in 和 fc_out）映射为内部使用的命名方式，例如 double_blocks.0.img_mlp.0.* → double_blocks.0.img_mlp.fc_in.*。
3. Double-stream QKV 融合：将 separate Q/K/V 投影（如 to_q、q_proj、query 等变体）映射到 fused qkv 张量，并指定合并索引（0、1、2）和合并数量（3）。
4. Double-stream out-projection 映射：将各种 out-projection 命名（to_out.0、to_add_out.0）统一映射到内部 proj 张量，并保留 weight/bias 后缀。
5. Q/K norm 别名与 weight→scale 转换：将 norm_q/norm_k 映射为 norm.query_norm / norm.key_norm，并将 HuggingFace 风格的 .weight 转换为内部使用的 .scale。
6. Single-stream 块处理：支持 single_transformer_blocks 和 single_blocks 两种导出命名，并将 attn 中的 Q/K/V 及 MLP 第一层（proj_mlp/mlp_fc1）打包到 linear1（merge_index 0-3），out-projection 映射到 linear2，norm 规则与 double-stream 类似。

对应的实现全部在 python/sglang/multimodal_gen/configs/models/dits/hunyuan3d.py 中完成。

关键文件：

- python/sglang/multimodal_gen/configs/models/dits/hunyuan3d.py（模块 参数映射；类别 source；类型 data-contract）：核心变更文件，扩展了 param_names_mapping 字典以支持 Hunyuan3D-2 DiT 官方 checkpoint 参数名映射。

关键符号：未识别

关键源码片段

python/sglang/multimodal_gen/configs/models/dits/hunyuan3d.py

核心变更文件，扩展了 param_names_mapping 字典以支持 Hunyuan3D-2 DiT 官方 checkpoint 参数名映射。

```
# SPDX-License-Identifier: Apache-2.0
from dataclasses import dataclass, field

from sglang.multimodal_gen.configs.models.dits.base import DiTArchConfig, DiTConfig


@dataclass
class Hunyuan3DDiTArchConfig(DiTArchConfig):
    """Architecture config for Hunyuan3D DiT (Flux-style for Hunyuan3D-2.0)."""

    param_names_mapping: dict = field(
        default_factory=lambda: {
            # 去除官方 safetensors 中可能存在的 "model." 前缀，使后续规则无需重复匹配
            r"^model\.(.*)$": r"\1",

            # Double-stream MLP 层重命名（索引 0 -> fc_in, 2 -> fc_out)
            r"^(double_blocks\.\d+\.img_mlp)\.0\.(.*)$": r"\1.fc_in.\2",
            r"^(double_blocks\.\d+\.img_mlp)\.2\.(.*)$": r"\1.fc_out.\2",
            r"^(double_blocks\.\d+\.txt_mlp)\.0\.(.*)$": r"\1.fc_in.\2",
            r"^(double_blocks\.\d+\.txt_mlp)\.2\.(.*)$": r"\1.fc_out.\2",

            # Double-stream attention: 将 split 的 Q/K/V 融合成 fused qkv, 并指定合并索引和数量
            r"^(double_blocks\.\d+\.(?:txt_attn|img_attn))\.(?:to_q|q_proj|query)\.(.*)$": (
                r"\1.qkv.\2", 0, 3),
            r"^(double_blocks\.\d+\.(?:txt_attn|img_attn))\.(?:to_k|k_proj|key)\.(.*)$": (
                r"\1.qkv.\2", 1, 3),
            r"^(double_blocks\.\d+\.(?:txt_attn|img_attn))\.(?:to_v|v_proj|value)\.(.*)$": (
                r"\1.qkv.\2", 2, 3),

            # Double-stream out-projection: 多种命名统一到 proj
            r"^(double_blocks\.\d+\.(?:txt_attn|img_attn))\.to_out(?:\.0)?\.(weight|bias)$": r"\1.proj.
            \2",
            r"^(double_blocks\.\d+\.txt_attn)\.to_add_out(?:\.0)?\.(weight|bias)$": r"\1.proj.\2",

            # Q/K norm 别名 + HF "weight" -> 内部 "scale"
```

```
            r"^(double_blocks\.\d+\.(?:txt_attn|img_attn))\.norm_q\.(.*)$": r"\1.norm.query_norm.\2",
            r"^(double_blocks\.\d+\.(?:txt_attn|img_attn))\.norm_k\.(.*)$": r"\1.norm.key_norm.\2",
            r"^(.*norm\.query_norm)\.weight$": r"\1.scale",
            r"^(.*norm\.key_norm)\.weight$": r"\1.scale",

            # Single-stream: 支持两种导出命名，QKV+MLP 打包进 linear1，out-proj 进 linear2
            r"^(?:single_blocks|single_transformer_blocks)\.(\d+)\.attn\.(?:to_q|q_proj|query)\.(.*)$":
            (
                r"single_blocks.\1.linear1.\2", 0, 4),
            r"^(?:single_blocks|single_transformer_blocks)\.(\d+)\.attn\.(?:to_k|k_proj|key)\.(.*)$": (
                r"single_blocks.\1.linear1.\2", 1, 4),
            r"^(?:single_blocks|single_transformer_blocks)\.(\d+)\.attn\.(?:to_v|v_proj|value)\.(.*)$":
            (
                r"single_blocks.\1.linear1.\2", 2, 4),
            r"^(?:single_blocks|single_transformer_blocks)\.(\d+)\.(?:proj_mlp|mlp_fc1)\.(.*)$": (
                r"single_blocks.\1.linear1.\2", 3, 4),
            r"^(?:single_blocks|single_transformer_blocks)\.(\d+)\.(?:proj_out|out_proj)(?:\.0)?\.
            (weight|bias)$": r"single_blocks.\1.linear2.\2",
            r"^(?:single_blocks|single_transformer_blocks)\.(\d+)\.attn\.to_out(?:\.0)?\.(weight|bias)
            $": r"single_blocks.\1.linear2.\2",
            # Single-stream norm 映射
            r"^(?:single_blocks|single_transformer_blocks)\.(\d+)\.attn\.norm_q\.(.*)$": r"single_
            blocks.\1.norm.query_norm.\2",
            r"^(?:single_blocks|single_transformer_blocks)\.(\d+)\.attn\.norm_k\.(.*)$": r"single_
            blocks.\1.norm.key_norm.\2",
        }
    )

    in_channels: int = 64
    hidden_size: int = 1024
    num_attention_heads: int = 16
    num_layers: int = 16
    num_single_layers: int = 32
    mlp_ratio: float = 4.0
    context_in_dim: int = 1536
    axes_dim: tuple[int, ...] = (64,)
    theta: int = 10000
    qkv_bias: bool = True
    guidance_embed: bool = False
    time_factor: float = 1000.0

    def __post_init__(self) -> None:
        if self.num_channels_latents == 0:
            self.num_channels_latents = self.in_channels
        super().__post_init__()


@dataclass
class Hunyuan3DDiTConfig(DiTConfig):
```

```
    """DiT configuration for Hunyuan3D shape generation (Flux-style)."""

    arch_config: Hunyuan3DDiTArchConfig = field(default_factory=Hunyuan3DDiTArchConfig)
    subfolder: str = "hunyuan3d-dit-v2-0"
```

评论区精华

无实质性讨论，PR 由 maintainer mickqian 直接批准。

- 暂无高价值评论线程

风险与影响

- 风险：
  1. 正则覆盖不足：若 checkpoint 中存在未考虑到的命名模式（如其他的 prefix 或 MLP 层索引），仍可能加载失败。
  2. 正则顺序冲突：多个规则可能匹配同一参数名，顺序敏感，后续维护时需小心。
  3. 无测试覆盖：未增加专门的参数映射测试，回归风险依赖人工验证。
  4. 影响范围局限：仅影响权重加载路径，不影响运行时推理，因此性能风险低。 - 影响：对依赖 Hunyuan3D-2 DiT 模型的用户，此修复使其能正常加载官方权重，解锁模型使用。对系统无性能影响，代码改动集中。对团队维护而言，映射规则需持续对齐官方导出的命名变化。 - 风险标记：缺少测试覆盖，正则冲突风险，仅支持特定 checkpoint 命名

关联脉络

- PR #22289 [Bugfix] multimodal_gen(hunyuan3d): honor config precisions for delight/paint: 同属 Hunyuan3D 模型族的 bugfix，但关注精度配置而非参数映射，属于该模型在不同方面的修复。