PR #6505 完整报告

verl-project/verl

[veomni, cfg] feat: add missing config fields to veomni.yaml

合并时间：2026-06-10 11:21

原文链接：http://prhub.com.cn/verl-project/verl/pull/6505

执行摘要

- 一句话：补全 VeOmni 引擎缺失配置与 Qwen3.5 算子支持
- 推荐动作：本 PR 属于 VeOmni 配置体系的规范性补全，建议普通用户阅读 final_report_markdown 中的配置变更摘要了解可用选项；参与 VeOmni 开发的工程师应关注 review 中关于 FSDP2-only 和算子选择约束的讨论，避免后续添加不兼容字段。

功能与动机

VeOmni 引擎的 YAML 配置落后于 Python dataclass 定义，导致用户无法通过配置文件控制某些训练选项（如断点续训路径、熵计算分块）。同时为支持 Qwen3.5 新架构（GatedDeltaNet 等）需要暴露新增算子的 kernel 选择入口。

实现拆解

1. 引擎配置类扩充（verl/workers/config/engine.py）：在 VeOmniEngineConfig 中补充 load_checkpoint_path、entropy_from_logits_with_chunking、entropy_checkpointing、basic_modules 字段，并移除 wrap_policy、offload_policy、reshard_after_forward、use_orig_params 等 FSDP1 专属废弃项；新增 rms_norm_gated_implementation、causal_conv1d_implementation、chunk_gated_delta_rule_implementation 三个算子选择字段。
2. 算子配置传递（verl/workers/engine/veomni/transformer_impl.py）：在 _build_model_optimizer 方法的 OpsImplementationConfig 构造中添加三个新算子字段，使其被 VeOmni 底层识别并应用。
3. 默认配置文件同步（verl/trainer/config/engine/veomni.yaml）：新增 15 行配置项，覆盖断点路径、熵控制、新算子默认值（eager）及 basic_modules 列表，并在注释中说明每个字段的适用场景与限制。
4. 生成配置模板更新（verl/trainer/config/_generated_ppo_veomni_trainer.yaml）：在 actor、ref policy、critic 三段中同步新增相同的配置项，确保 V1 trainer 在 omegaconf 合并时不会出现未定义键错误。

关键文件：

- verl/workers/config/engine.py（模块 配置；类别 source；类型 core-logic；符号 VeOmniEngineConfig）：核心配置类修改，新增 / 移除字段以匹配 VeOmni v0.1.11 特性，定义新算子实现选择枚举。

- verl/workers/engine/veomni/transformer_impl.py（模块 引擎；类别 source；类型 core-logic；符号 VeOmniEngine._build_model_optimizer）：VeOmni 引擎实现，修改 _build_model_optimizer 以传递新算子配置到下层 OpsImplementationConfig。
- verl/trainer/config/engine/veomni.yaml（模块 配置；类别 config；类型 configuration）：VeOmni 引擎默认配置文件，新增 15 项配置，用户可直接参考或修改。
- verl/trainer/config/_generated_ppo_veomni_trainer.yaml（模块 配置；类别 config；类型 configuration）：自动生成的 PPO VeOmni 训练配置模板，同步新增字段以保持一致性。

关键符号：VeOmniEngine._build_model_optimizer

关键源码片段

verl/workers/config/engine.py

核心配置类修改，新增 / 移除字段以匹配 VeOmni v0.1.11 特性，定义新算子实现选择枚举。

```
# verl/workers/config/engine.py (partial, VeOmniEngineConfig 类定义摘录 )
@dataclass
class VeOmniEngineConfig(EngineConfig):
    """Configuration for VeOmni.

    The inheritance from BaseConfig provides omegaconf.DictConfig-like interface
    for a dataclass config.
    """

    # --- 基础并行与优化选项 ---
    param_offload: bool = False
    optimizer_offload: bool = False
    fsdp_size: int = -1
    ulysses_parallel_size: int = 1
    expert_parallel_size: int = 1
    init_device: str = "meta"
    enable_full_shard: bool = True
    enable_fsdp_offload: bool = False
    enable_reentrant: bool = False
    forward_prefetch: bool = True

    # --- 新增 / 保留字段（与 FSDP 通用引擎一致） ---
    # load_checkpoint_path: 断点续训路径，默认 None 不从 checkpoint 恢复
    load_checkpoint_path: Optional[str] = None
    entropy_from_logits_with_chunking: bool = False
    entropy_checkpointing: bool = False
    use_torch_compile: bool = False
    forward_only: bool = False
    basic_modules: list[str] = field(default_factory=list)
    # ... 其它基类字段

    # --- Qwen3.5 新增算子实现选择 ---
    # 仅在 VeOmni 使用 apply_ops_config 时生效，NPU 不支持非 eager 值
```

```
    rms_norm_gated_implementation: str = "eager"
    causal_conv1d_implementation: str = "eager"
    chunk_gated_delta_rule_implementation: str = "eager"

    # --- 注意：已移除的 FSDP1 字段不存在于此 class 中 ---
    #  (wrap_policy, offload_policy, reshard_after_forward, use_orig_params 等已删除)

    _mutable_fields = EngineConfig._mutable_fields | {"ulysses_sequence_parallel_size"}
```

verl/workers/engine/veomni/transformer_impl.py

VeOmni 引擎实现，修改 _build_model_optimizer 以传递新算子配置到下层 OpsImplementationConfig。

```
# verl/workers/engine/veomni/transformer_impl.py ( 部分 )
    def _build_model_optimizer(self):
        # 构造 OpsImplementationConfig，传递给 VeOmni 的 apply_per_model_patches
        ops_implementation = OpsImplementationConfig(
            attn_implementation=self.engine_config.attn_implementation,
            moe_implementation=self.engine_config.moe_implementation,
            cross_entropy_loss_implementation=self.engine_config.cross_entropy_loss_
            implementation,
            rms_norm_implementation=self.engine_config.rms_norm_implementation,
            swiglu_mlp_implementation=self.engine_config.swiglu_mlp_implementation,
            rotary_pos_emb_implementation=self.engine_config.rotary_pos_emb_implementation,
            load_balancing_loss_implementation=self.engine_config.load_balancing_loss_
            implementation,
            # 以下为新增的 Qwen3.5 算子字段
            rms_norm_gated_implementation=self.engine_config.rms_norm_gated_implementation,
            causal_conv1d_implementation=self.engine_config.causal_conv1d_implementation,
            chunk_gated_delta_rule_implementation=self.engine_config.chunk_gated_delta_rule_
            implementation,
        )
        # 后续使用 ops_implementation 构建模型
        module = build_foundation_model(
            ...,
            ops_implementation=ops_implementation,
        )
```

verl/trainer/config/engine/veomni.yaml

VeOmni 引擎默认配置文件，新增 15 项配置，用户可直接参考或修改。

```
# verl/trainer/config/engine/veomni.yaml ( 部分新字段 )
# Path to load checkpoint from, if any
load_checkpoint_path: null

# Whether to use entropy_from_logits_with_chunking in fsdp.
entropy_from_logits_with_chunking: false

# Whether to use entropy checkpointing in fsdp.
```

```
entropy_checkpointing: false

# ... ( 中间略 )

# Kernel backends for Qwen3.5 components (all default to eager).
# - rms_norm_gated: eager (HF Qwen3_5RMSNormGated) / fla (GPU only)
# - causal_conv1d: eager (no varlen support) / fla (GPU only)
# - chunk_gated_delta_rule: eager (no cu_seqlens) / fla (GPU) / flash_qla (Hopper only)
# NPU 用户请保持 eager，否则运行时报错。
rms_norm_gated_implementation: eager
causal_conv1d_implementation: eager
chunk_gated_delta_rule_implementation: eager

# List of basic modules to use (e.g. added by device_patches)
basic_modules: []
```

评论区精华

gemini-code-assist[bot] 提出多项高优先级问题：

- grad_offload、offload_policy、wrap_policy 参数在 VeOmni 中不被支持（VeOmni 使用 FSDP2，不支持独立梯度卸载或 FSDP1 的 wrap policy）。
- dtype 配置无效：当前 _build_model_optimizer 硬编码 torch_dtype 为 "float32" if mixed_precision else "bfloat16"，不会读取 dtype 字段。
- use_orig_params 是 FSDP1 专用，VeOmni 固定使用 FSDP2，应移除。

FoolPlayer（维护者）确认上述问题并补充：VeOmni 自 v0.1.10 起已移除 FSDP1 支持，即将随 verl 0.8.0 使用的 v0.1.11 完全只支持 FSDP2；对于 model dtype，建议同时添加断言确保 bfloat16 时 mixed_precision=True 一致。

作者后续提交清理：在后续 commit 中删除了 offload_policy、grad_offload、wrap_policy、dtype、use_orig_params 等冗余项，并重命名 offload_policy 为 offload_policy（但最终确认删除）。最终通过 review 与 CI。

- VeOmni 配置冗余项清理 (design): 作者在后续 commit 中移除了这些冗余字段，仅保留与 VeOmni 兼容的配置。
- dtype 与 mixed_precision 交互 (correctness): 作者最终移除了 dtype 字段，保持原有 hardcode 逻辑，避免不一致。
- 新算子选择在 NPU 上的可用性 (question): 未添加主动校验，依赖用户阅读注释。保留为后续改进空间。

风险与影响

- 风险：
  1. 配置向前兼容：新增字段在不使用时保持默认值，不影响现有训练脚本；但若用户 YAML 中未更新到新默认值，且遗忘了某个必需字段，可能引发 omegaconf 合并错误（如 basic_modules 非空时与 _no_split_modules 合并行为可能变化）。

2. 新算子选择误用：部分算子实现（如 rms_norm_gated 的 fla、causal_conv1d 的 fla）在 NPU 上不可用，但当前注释仅提示未绑定；若用户设置非 *eager* 值在 NPU 运行则可能运行时异常，缺少主动校验。
3. 与 mixed_precision 交互：entropy_from_logits_with_chunking 与 mixed_precision 的交互未在代码中约束，用户可能配置不一致导致数值精度问题。
4. 测试覆盖：本次变更未新增测试，若底层 VeOmni 升级改变算子参数含义，会导致配置偏差。 - 影响：对用户：VeOmni 训练脚本编写者可利用新配置项控制断点续训、熵计算、Qwen3.5 特定算子后端，获得更细粒度的训练调优能力。移除已废弃 FSDP1 参数避免混淆。对系统：改动限于配置层与 VeOmni 引擎初始化路径，无运行时性能影响。CI 验证通过，日志显示配置生效。对团队：统一了 config 入口与 Python 定义，降低后续维护成本。新增算子选择字段为后续 Qwen3.5 支持奠定配置基础。

- 风险标记：配置项与实现脱钩风险，NPU 算子选择缺少校验，缺少配套测试

关联脉络

- PR #6853 [trainer, rollout] feat: log rollout moe load-balance metrics: 同样修改了 verl/workers/engine/veomni/transformer_impl.py，涉及 VeOmni 引擎内监控相关变更，与本 PR 的配置扩展属于同一引擎功能线。