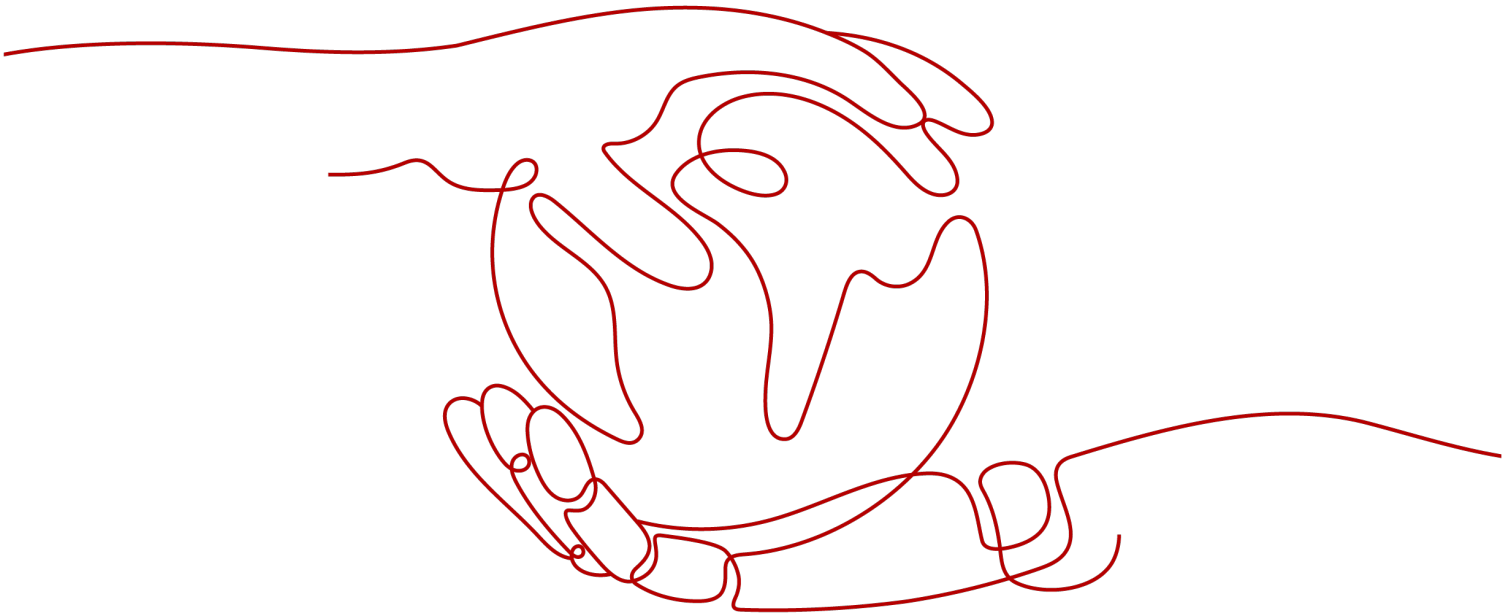
ModelArts

服务公告

文档版本	01
发布日期	2025-02-19



华为云计算技术有限公司

商标声明


和其他华为商标均为华为技术有限公司的商标。
本文档提及的其他所有商标或注册商标，由各自的所有人拥有。

注意

您购买的产品、服务或特性等应受华为云计算技术有限公司商业合同和条款的约束，本文档中描述的全部或部分产品、服务或特性可能不在您的购买或使用范围之内。除非合同另有约定，华为云计算技术有限公司对本文档内容不做任何明示或暗示的声明或保证。

由于产品版本升级或其他原因，本文档内容会不定期进行更新。除非另有约定，本文档仅作为使用指导，本文档中的所有陈述、信息和建议不构成任何明示或暗示的担保。

华为云计算技术有限公司

地址： 贵州省贵安新区黔中大道交兴功路华为云数据中心 邮编：550029

网址： https://www.huaweicloud.com/

目 录

1 下线公告

1.1 【下线公告】华为云 ModelArts 自动学习下线公告

华为云计划于2025/05/23 00:00（北京时间）将AI开发平台ModelArts自动学习模块正式下线。

下线范围

下线区域：华为云全部Region

下线影响

正式下线后，所有用户将无法使用自动学习模块创建项目，但仍可在Workflow模块查看、使用历史创建的自动学习作业。

如您有任何问题，可随时通过工单或者服务热线（+86-4000-955-988或+86-950808）与我们联系。

1.2 【下线公告】华为云 ModelArts 自动学习模块的文本分类功能下线公告

华为云计划于2024/12/06 00:00（北京时间）将AI开发平台ModelArts自动学习模块的文本分类功能正式下线。

下线范围

下线Region：华为云全部Region。

下线影响

ModelArts自动学习-文本分类正式下线后，所有用户将无法使用自动学习的文本分类功能创建项目，但仍可查看历史使用文本分类功能创建的作业。

如您有任何问题，可随时通过工单或者服务热线（+86-4000-955-988或+86-950808）与我们联系。

感谢您对华为云的支持！

1.3 【下线公告】华为云 ModelArts 服务旧版数据集下线公告

华为云计划于2024/10/31 00:00（北京时间）用AI开发平台ModelArts的新版数据集全面替代旧版数据集，旧版数据集正式下线。

下线范围

下线区域：华北-北京四（其他区域已下线）

受影响服务

ModelArts旧版数据集。

下线影响

正式下线后，所有用户将无法使用旧版数据集。为了避免影响您的业务，建议您在2024/10/30 23:59:59（北京时间）前备份数据或切换至新版数据集。

如您有任何问题，可随时通过工单或者服务热线（+86-4000-955-988或+86-950808）与我们联系。

1.4 【下线公告】华为云 ModelArts 服务模型转换下线公告

华为云ModelArts服务模型转换在2024年4月30日 00:00(北京时间)正式下线。

下线范围

下线区域：华为云全部Region

下线影响

正式下线后，用户将无法再使用模型转换的功能，包括创建和删除模型转换任务、查询模型转换任务列表和详情功能。

如您有任何问题，可随时通过**工单**或者服务热线（+86-4000-955-988或+86-950808）与我们联系。

常见问题

为什么要下线模型转换?

ModelArts模型转换向AI开发者提供了便捷的模型转换页面，将Tensorflow和Caffe框架的模型格式转换为MindSpore的模型格式，即模型后缀为.om，使之能在昇腾硬件中进行推理。由于产品演进规划，后续昇腾硬件推理时主要使用后缀为.mindir的模型格式，因此ModelArts下线.om格式的模型转换能力，在ModelArts中逐步增加.mindir格式的支持能力。

下线模型转换后是否有替代功能?

您可以通过链接下载**ATC模型转换工具**，按照指导，在线下转换成.om格式模型。

ModelArts中是否还会增加模型转换的能力?

ModelArts开发环境中在贵阳一Region，支持将ONNX或PyTorch模型转换到.mindir格式。其它能力在持续增加中。如果您暂时无法在该region中使用该能力，您可以通过链接下载**MindSpore Lite离线转换模型工具**，线下将其转换为.mindir格式模型。

1.5 【下线公告】华为云 ModelArts MindStudio/ML Studio/ModelBox 镜像下线公告

华为云ModelArts服务MindStudio，ML Studio，ModelBox镜像将在2024年6月30日00:00（北京时间）正式退市。

下线范围

下线Region：华为云全部Region

下线影响

正式下线后，ModelArts Notebook创建页面将不再呈现MindStudio、ML Studio、ModelBox这几个镜像，将无法使用这几个镜像新建Notebook实例，用户已有实例仍可以继续使用。后续删除实例后将无法再新建。如您有任何问题，可随时通过工单或者服务热线（4000-955-988或950808）与我们联系。

常见问题

下线镜像对现有用户的使用是否有影响?

下线镜像对已有用户不影响，用户可以继续使用已有实例启动Notebook，但是需要注意删除实例后无法再新建实例。

镜像下线后是否可以继续基于该镜像新建实例?

镜像下线后无法使用该镜像新建实例，界面不会呈现了。

镜像下线后用户还想继续使用，怎么办?

如果想长期使用该镜像，建议用户在镜像下线前保存自定义镜像使用，镜像下线后不会影响自定义镜像使用。

1.6 【下线公告】华为云 ModelArts 算法套件下线公告

华为云ModelArts服务算法套件将在2024年6月30日00:00（北京时间）正式退市。

下线范围

下线Region：华为云全部Region。

下线影响

正式下线后，ModelArts Notebook中将不会预置算法套件相关工具ma-cau和ma-cau-adapter，ma-cli命令将不支持创建算法工程，无法在Notebook中基于已有算法工程进行资产（数据、模型权重、算法文件）安装、模型开发、训练和推理部署等任务。

如您有任何问题，可随时通过**工单**或者服务热线（4000-955-988或950808）与我们联系。

1.7 【下线公告】华为云 ModelArts 服务旧版训练管理下线公告

华为云ModelArts服务旧版训练管理在2023年6月30日 00:00(北京时间)正式退市。

下线范围

下线区域：华为云全部Region

下线影响

正式下线后，用户将无法再使用旧版训练管理的功能，包括旧版训练作业、训练参数管理、可视化作业功能，建议将相关作业迁移到新版训练管理。

如您有任何问题，可随时通过**工单**或者服务热线（4000-955-988或950808）与我们联系。

常见问题

为什么要下线旧版训练管理?

ModelArts旧版训练全面上线以后为众多开发者提供了AI训练能力，其中训练服务作为基础服务之一，经过持续迭代已经无法完全满足众多开发者的新特性需求。基于服务演进，ModelArts团队已于2021年上线新版训练，力求解决存在的历史问题，并为新特性提供高性能、高易用、可扩展、可演进的底座，给用户提供更好的AI训练体验，打造易用、高效的AI平台。

下线旧版训练管理对现有用户的使用是否有影响?

正在使用的训练作业不受影响，但是用户无法使用旧版训练创建新的作业。

旧版训练管理是否停止新购?

是的，旧版训练管理将于2023年6月30日 00:00(北京时间)正式退市。

旧版训练管理如何升级到新版训练?

请参考新版训练指导文档（**模型训练**）来体验新版训练。

旧版训练迁移至新版训练需要注意哪些问题?

新版训练和旧版训练的差异主要体现在以下3点。

- 新旧版创建训练作业方式差异
- 新旧版训练代码适配的差异
- 新旧版训练预置引擎差异

2 产品发布说明

2.1 ModelArts 版本配套关系表

当前华为云中国站和国际站所有Region均已上线ModelArts 6.7.0版本。 ModelArts 6.7.0版本中针对Ascend Snt9B资源的周边依赖组件配套版本关系如下表所示。

表 2-1 ModelArts 6.7.0 版本配套关系表

强依赖组件	Ascend Snt9B配套版本
CCE	1.28（推荐）/1.25/1.23（存量）
Volcano插件	1.15.8
ModelArts Device-Plugin	1.1.0
huawei-npu	2.1.22
Lite模式DevServer节点操作系统	HCE2.0（推荐）/EulerOS 2.10
Lite模式Cluster节点操作系统	EulerOS 2.10（CCE标准版）/HCE2.0（CCE Turbo）
Standard模式集群节点操作系统	EulerOS 2.10（CCE标准版）
NPU固件&驱动	7.1.0.9.220-23.0.6（推荐） 7.3.0.1.231-24.1.rc2（白名单）
NPU CANN	8.0.RC2（推荐） 8.0.RC1 7.0.1.1
NPU MindSpore	2.3.0（推荐） 2.2.12 2.2.10
NPU PyTorch	2.1.0（推荐）

强依赖组件	Ascend Snt9B配套版本
预置统一镜像	pytorch_2.1.0-cann_8.0.rc2-py_3.9-euler_2.10.7-aarch64-snt9b mindspore_2.3.0-cann_8.0.rc2-py_3.9-euler_2.10.7-aarch64-snt9b pytorch_1.11.0-cann_8.0.rc2-py_3.9-euler_2.10.7-aarch64-snt9b
SFS Turbo Client+	23.09.03
AI Turbo SDK	23.12.3

2.2 昇腾云服务 6.3.912 版本说明

本文档主要介绍昇腾云服务6.3.912版本配套的镜像地址、软件包获取方式和支持的特性能力。

当前版本仅适用于华为公有云。

配套的基础镜像

芯片	镜像地址	获取方式	镜像软件说明
Snt9B	**PyTorch2.1.0**： swr.cn-southwest-2.myhuaweicloud.com/atelier/pytorch_2_1_ascend:pytorch_2.1.0-cann_8.0.rc3-py_3.9-hce_2.0.2409-aarch64-snt9b-20241213131522-aafe527 **PyTorch2.3.1**： swr.cn-southwest-2.myhuaweicloud.com/atelier/pytorch_2_3_ascend:pytorch_2.3.1-cann_8.0.rc3-py_3.10-hce_2.0.2409-aarch64-snt9b-20241213131522-aafe527 **MindSpore**： swr.cn-southwest-2.myhuaweicloud.com/atelier/mindspore_2_4_ascend:mindspore_2.4.0-cann_8.0.rc3-py_3.9-hce_2.0.2409-aarch64-snt9b-20241113174059-fcd3700	镜像发布到SWR， region：西南-贵阳一， 从SWR拉取	固件驱动：23.0.6 CANN：cann_8.0.rc3 容器镜像OS：hce_2.0 PyTorch：pytorch_2.1.0、pytorch2.3.1 MindSpore：MindSpore 2.4.0 FrameworkPTAdapter：6.0.RC3 CCE：如果用到CCE，版本要求是CCE Turbo v1.28及以上

芯片	镜像地址	获取方式	镜像软件说明
300iDUO	**PyTorch**： swr.cn-southwest-2.myhuaweicloud.com/atelier/pytorch_2_1_ascend:pytorch_2.1.0-cann_8.0.rc3-py_3.9-hce_2.0.2406-aarch64-snt3p-20240906180137-154bd1b	镜像发布到SWR， region：西南-贵阳一， 从SWR拉取	固件驱动:24.1.rc2.3 CANN：cann_8.0.rc3 容器镜像OS：hce_2.0 PyTorch：pytorch_2.1.0 MindSpore lite：2.3.0 FrameworkPTAdapter：6.0.RC3

软件包获取地址

软件包名称	软件包说明	获取地址
AscendCloud-6.3.912-xxx.zip	包含 1. 三方大模型训练和推理代码包：AscendCloud-LLM 2. AIGC代码包：AscendCloud-AIGC 3. CV代码包：AscendCloud-CV 4. 算子依赖包：AscendCloud-OPP	获取路径：**Support-E**，在此路径中查找下载ModelArts 6.3.912 版本。 **说明** 如果上述软件获取路径打开后未显示相应的软件信息，说明您没有下载权限，请联系您所在企业的华为方技术支持下载获取。

支持的特性

表 2-2 本版本支持的特性说明

分类	软件包特性说明	参考文档
三方大模型，包名：AscendCloud-LLM	支持如下模型适配PyTorch-NPU的训练(ModelLink) 1. llama2-7b 2. llama2-13b 3. llama2-70b 4. qwen-7b 5. qwen-14b 6. qwen-72b 7. baichuan2-13b 8. chatglm3-6b 9. llama3-8b 10.llama3-70b 11.yi-6B 12.yi-34B 13.qwen1.5-7B 14.qwen1.5-14B 15.qwen1.5-32B 16.qwen1.5-72B 17.qwen2-0.5b 18.qwen2-1.5b 19.qwen2-7b 20.qwen2-72b 21.glm4-9b 22.mistral-7b 23.llama3.1-8b 24.llama3.1-70b 25.qwen2.5-0.5b 26.qwen2.5-7b 27.qwen2.5-14b 28.qwen2.5-32b 29.qwen2.5-72b 30.llama3.2-1b 31.llama3.2-3b	**LLM开源大模型基于DevServer适配ModelLinkPyTorch NPU训练指导** **LLM开源大模型基于DevServer适配LLamaFactory PyTorch NPU训练指导** **LLM开源大模型基于Standard+OBS适配PyTorch NPU训练指导** **LLM开源大模型基于Standard+OBS+SFS适配PyTorch NPU训练指导** **LLM开源大模型基于Lite Cluster适配PyTorch NPU训练指导**

分类	软件包特性说明	参考文档
	支持如下模型适配PyTorch-NPU的训练(LlamaFactory) 1. llama2-7b 2. llama2-13b 3. llama2-70b 4. llama3-8b 5. llama3-70b 6. llama3.1-8b 7. llama3.1-70b 8. qwen1.5-7b 9. qwen1.5-14b 10.qwen1.5-32b 11.qwen1.5-72b 12.yi-6b 13.yi-34b 14.qwen2-0.5b 15.qwen2-1.5b 16.qwen2-7b 17.qwen2-72b 18.qwen2_vl-2b 19.qwen2_vl-7b 20.qwen2_vl-72b 21.falcon-11B 22.glm4-9b 23.qwen2.5-0.5b 24.qwen2.5-7b 25.qwen2.5-14b 26.qwen2.5-32b 27.qwen2.5-72b 28.llama3.2-1b 29.llama3.2-3b	

分类	软件包特性说明	参考文档
	支持如下模型适配PyTorch-NPU的推理(Ascend-vLLM框架)： 1. llama-7B 2. llama-13b 3. llama-65b 4. llama2-7b 5. llama2-13b 6. llama2-70b 7. llama3-8b 8. llama3-70b 9. yi-6b 10.yi-9b 11.yi-34b 12.deepseek-llm-7b 13.deepseek-coder-instruct-33b 14.deepseek-llm-67b 15.qwen-7b 16.qwen-14b 17.qwen-72b 18.qwen1.5-0.5b 19.qwen1.5-7b 20.qwen1.5-1.8b 21.qwen1.5-14b 22.qwen1.5-32b 23.qwen1.5-72b 24.qwen1.5-110b 25.qwen2-0.5b 26.qwen2-1.5b 27.qwen2-7b 28.qwen2-72b 29.qwen2.5-0.5b 30.qwen2.5-1.5b 31.qwen2.5-3b 32.qwen2.5-7b 33.qwen2.5-14b 34.qwen2.5-32b 35.qwen2.5-72b	**LLM开源大模型基于Lite Server适配PyTorch NPU推理指导** **LLM开源大模型基于Standard适配PyTorch NPU推理指导** **LLM开源大模型基于Lite Cluster适配PyTorch NPU推理指导**

分类	软件包特性说明	参考文档
	36.baichuan2-7b 37.baichuan2-13b 38.chatglm2-6b 39.chatglm3-6b 40.glm-4-9b 41.gemma-2b 42.gemma-7b 43.mistral-7b 44.mixtral 8*7B 45.falcon2-11b 46.qwen2-57b-a14b 47.llama3.1-8b 48.llama3.1-70b 49.llama-3.1-405B 50.llama-3.2-1B 51.llama-3.2-3B 52.llava-1.5-7b 53.llava-1.5-13b 54.llava-v1.6-7b 55.llava-v1.6-13b 56.llava-v1.6-34b 57.internvl2-8B 58.internvl2-26B 59.internvl2-40B 60.internVL2-Llama3-76B 61.MiniCPM-v2.6 62.deepseek-v2-236B 63.deepseek-coder-v2-lite-16B 64.qwen2-vl-2B 65.qwen2-vl-7B 66.qwen2-vl-72B 67.qwen-vl 68.qwen-vl-chat 69.MiniCPM-v2 70.gte-Qwen2-7B-instruct 71.llava-onevision-qwen2-0.5b-ov-hf	

分类	软件包特性说明	参考文档
	72.llava-onevision-qwen2-7b-ov-hf Ascend-vllm支持如下推理特性： 1. 支持分离部署 2. 支持多机推理 3. 支持大小模型投机推理及eagle投机推理 4. 支持chunked prefill特性 5. 支持automatic prefix caching 6. 支持multi-lora特性 7. 支持W4A16、W8A16和W8A8量化 8. 升级vLLM 0.6.3 9. 支持流水线并行 说明：具体模型支持的特性请参见大模型推理指导文档	

分类	软件包特性说明	参考文档
AIGC，包名：AscendCloud-AIGC	支持如下框架或模型基于PyTorch NPU推理（PyTorch框架）： 1. ComfyUI 2. Diffusers 3. Wav2Lip 4. OpenSora1.2 5. OpenSoraPlan1.0 6. FLUX.1 7. Hunyuan-Dit 8. Qwen-VL 9. CogVideoX 10.LLama-VID 11.MiniCPM-V2.0 12.SD3 13.SD3.5 支持如下框架或模型基于PyTorch NPU的训练（PyTorch框架）： 1. Qwen-VL 2. Diffusers 3. Koyha_ss 4. Wav2Lip 5. InternVL2 6. OpenSora1.2 7. OpenSoraPlan1.0 8. CogVideoX 9. LLaVA-NeXT 10.LLaVA 11.MiniCPM-V2.0 12.FLUX.1 13.Llmma-3.2-11b 14.CogVideoX1.5 5b 15.MiniCPM-V2.6	**文生图模型训练推理** **文生视频模型训练推理** **多模态模型训练推理** **数字人模型训练推理**

分类	软件包特性说明	参考文档
CV，包名：AscendCloud-CV	支持如下模型适配MindSpore Lite的推理： 1. Yolov8 2. Bert 支持如下模型适配PyTorch NPU的推理： 1. Paraformer	内容审核模型推理
算子，包名：AscendCloud-OPP	1. Scatter、Gather算子性能提升，满足MoE训练场景 2. matmul、swiglu、rope等算子性能提升，支持vllm推理场景 3. 支持random随机数算子，优化FFN算子，满足AIGC等场景 4. 支持自定义交叉熵融合算子，满足BMTrain框架训练性能要求 5. 优化PageAttention算子，满足vllm投机推理场景 6. 支持CopyBlocks算子，满足vllm框架beam search解码场景 7. 支持AdvanceStep算子，满足vllm投机推理场景 8. 多个融合算子支持PTA图模式适配，满足AIGC场景 9. 支持两种版本配套算子包（torch2.1.0和python3.9、torch2.3.1和python3.10）	无

2.3 昇腾云服务 6.3.911 版本说明

本文档主要介绍昇腾云服务6.3.911版本配套的镜像地址、软件包获取方式和支持的特性能力。

当前版本仅适用于华为公有云。

配套的基础镜像

芯片	镜像地址	获取方式	镜像软件说明
Snt9B	**PyTorch2.1.0**： swr.cn-southwest-2.myhuaweicloud.com/atelier/pytorch_2_1_ascend:pytorch_2.1.0-cann_8.0.rc3-py_3.9-hce_2.0.2409-aarch64-snt9b-20241112192643-c45ac6b **PyTorch2.3.1**： swr.cn-southwest-2.myhuaweicloud.com/atelier/pytorch_2_3_ascend:pytorch_2.3.1-cann_8.0.rc3-py_3.10-hce_2.0.2409-aarch64-snt9b-20241114095658-d7e26d8 **MindSpore**： swr.cn-southwest-2.myhuaweicloud.com/atelier/mindspore_2_4_ascend:mindspore_2.4.0-cann_8.0.rc3-py_3.9-hce_2.0.2409-aarch64-snt9b-20241113174059-fcd3700	镜像发布到SWR， region：西南-贵阳一， 从SWR拉取	固件驱动：23.0.6 CANN：cann_8.0.rc3 容器镜像OS：hce_2.0 PyTorch：pytorch_2.1.0、pytorch2.3.1 MindSpore：MindSpore 2.4.0 FrameworkPTAdapter：6.0.RC3 CCE：如果用到CCE，版本要求是CCE Turbo v1.28及以上
300iDUO	**PyTorch**： swr.cn-southwest-2.myhuaweicloud.com/atelier/pytorch_2_1_ascend:pytorch_2.1.0-cann_8.0.rc3-py_3.9-hce_2.0.2406-aarch64-snt3p-20240906180137-154bd1b	镜像发布到SWR， region：西南-贵阳一， 从SWR拉取	固件驱动:24.1.rc2.3 CANN：cann_8.0.rc3 容器镜像OS：hce_2.0 PyTorch：pytorch_2.1.0 MindSpore lite：2.3.0 FrameworkPTAdapter：6.0.RC3

软件包获取地址

软件包名称	软件包说明	获取地址
AscendCloud-6.3.911-xxx.zip	包含 1. 三方大模型训练和推理代码包：AscendCloud-LLM 2. AIGC代码包：AscendCloud-AIGC 3. CV代码包：AscendCloud-CV 4. 算子依赖包：AscendCloud-OPP	获取路径：**Support-E**，在此路径中查找下载ModelArts 6.3.911 版本。 **说明** 如果上述软件获取路径打开后未显示相应的软件信息，说明您没有下载权限，请联系您所在企业的华为方技术支持下载获取。

支持的特性

表 2-3 本版本支持的特性说明

分类	软件包特性说明	参考文档
三方大模型，包名：AscendCloud-LLM	支持如下模型适配PyTorch-NPU的训练(ModelLink) 1. llama2-7b 2. llama2-13b 3. llama2-70b 4. qwen-7b 5. qwen-14b 6. qwen-72b 7. baichuan2-13b 8. chatglm3-6b 9. llama3-8b 10.llama3-70b 11.yi-6B 12.yi-34B 13.qwen1.5-7B 14.qwen1.5-14B 15.qwen1.5-32B 16.qwen1.5-72B 17.qwen2-0.5b 18.qwen2-1.5b 19.qwen2-7b 20.qwen2-72b 21.glm4-9b 22.mistral-7b 23.mixtral-8x7b 24.llama3.1-8b 25.llama3.1-70b 26.qwen2.5-0.5b 27.qwen2.5-7b 28.qwen2.5-14b 29.qwen2.5-32b 30.qwen2.5-72b 31.llama3.2-1b 32.llama3.2-3b 支持如下模型适配PyTorch-NPU的训练(LlamaFactory)	**LLM开源大模型基于DevServer适配ModelLinkPyTorch NPU训练指导** **LLM开源大模型基于DevServer适配LLamaFactory PyTorch NPU训练指导** **LLM开源大模型基于Standard+OBS适配PyTorch NPU训练指导** **LLM开源大模型基于Standard+OBS+SFS适配PyTorch NPU训练指导** **LLM开源大模型基于Lite Cluster适配PyTorch NPU训练指导**

分类	软件包特性说明	参考文档
	1. llama2-7b 2. llama2-13b 3. llama2-70b 4. llama3-8b 5. llama3-70b 6. llama3.1-8b 7. llama3.1-70b 8. qwen1.5-7b 9. qwen1.5-14b 10.qwen1.5-32b 11.qwen1.5-72b 12.yi-6b 13.yi-34b 14.qwen2-0.5b 15.qwen2-1.5b 16.qwen2-7b 17.qwen2-72b 18.qwen2_vl-2b 19.qwen2_vl-7b 20.falcon-11B 21.glm4-9b 22.qwen2.5-0.5b 23.qwen2.5-7b 24.qwen2.5-14b 25.qwen2.5-32b 26.qwen2.5-72b 27.llama3.2-1b 28.llama3.2-3b	

分类	软件包特性说明	参考文档
	支持如下模型适配PyTorch-NPU的推理(Ascend-vLLM框架)： 1. llama-7B 2. llama-13b 3. llama-65b 4. llama2-7b 5. llama2-13b 6. llama2-70b 7. llama3-8b 8. llama3-70b 9. yi-6b 10.yi-9b 11.yi-34b 12.deepseek-llm-7b 13.deepseek-coder-instruct-33b 14.deepseek-llm-67b 15.qwen-7b 16.qwen-14b 17.qwen-72b 18.qwen1.5-0.5b 19.qwen1.5-7b 20.qwen1.5-1.8b 21.qwen1.5-14b 22.qwen1.5-32b 23.qwen1.5-72b 24.qwen1.5-110b 25.qwen2-0.5b 26.qwen2-1.5b 27.qwen2-7b 28.qwen2-72b 29.qwen2.5-0.5b 30.qwen2.5-1.5b 31.qwen2.5-3b 32.qwen2.5-7b 33.qwen2.5-14b 34.qwen2.5-32b 35.qwen2.5-72b	**LLM开源大模型基于Lite Server适配PyTorch NPU推理指导** **LLM开源大模型基于Standard适配PyTorch NPU推理指导** **LLM开源大模型基于Lite Cluster适配PyTorch NPU推理指导**

分类	软件包特性说明	参考文档
	36.baichuan2-7b 37.baichuan2-13b 38.chatglm2-6b 39.chatglm3-6b 40.glm-4-9b 41.gemma-2b 42.gemma-7b 43.mistral-7b 44.mixtral 8*7B 45.falcon2-11b 46.qwen2-57b-a14b 47.llama3.1-8b 48.llama3.1-70b 49.llama-3.1-405B 50.llama-3.2-1B 51.llama-3.2-3B 52.llava-1.5-7b 53.llava-1.5-13b 54.llava-v1.6-7b 55.llava-v1.6-13b 56.llava-v1.6-34b 57.internvl2-8B 58.internvl2-26B 59.internvl2-40B 60.internVL2-Llama3-76B 61.MiniCPM-v2.6 62.deepseek-v2-236B 63.deepseek-coder-v2-lite-16B 64.qwen2-vl-2B 65.qwen2-vl-7B 66.qwen2-vl-72B 67.qwen-vl 68.qwen-vl-chat 69.MiniCPM-v2 Ascend-vllm支持如下推理特性： 1. 支持分离部署 2. 支持多机推理	

分类	软件包特性说明	参考文档
	3. 支持大小模型投机推理及eagle投机推理 4. 支持chunked prefill特性 5. 支持automatic prefix caching 6. 支持multi-lora特性 7. 支持W4A16、W8A16和W8A8量化 8. 升级vLLM 0.6.3 说明：具体模型支持的特性请参见大模型推理指导文档	

分类	软件包特性说明	参考文档
AIGC，包名：AscendCloud-AIGC	支持如下框架或模型基于DevServer的PyTorch NPU推理（PyTorch框架）： 1. ComfyUI 2. Diffusers 3. Stable-diffusion-webui 4. Wav2Lip 5. OpenSora1.2 6. OpenSoraPlan1.0 7. MiniCPM-V2.6 8. FLUX.1 9. Hunyuan-Dit 10.Qwen-VL 11.CogVideoX 12.LLama-VID 13.MiniCPM-V2.0 支持如下框架或模型基于DevServer的PyTorch NPU的训练（PyTorch框架）： 1. Qwen-VL 2. Diffusers 3. Koyha_ss 4. Wav2Lip 5. InternVL2 6. OpenSora1.2 7. OpenSoraPlan1.0 8. CogVideoX 9. LLaVA-NeXT 10.LLaVA 11.MiniCPM-V2.0 12.FLUX.1 13.Llmma-3.2-11b	**文生图模型训练推理** **文生视频模型训练推理** **多模态模型训练推理** **数字人模型训练推理**
CV，包名：AscendCloud-CV	支持如下模型适配MindSpore Lite的推理： 1. Yolov8 2. Bert 支持如下模型适配PyTorch NPU的推理： 1. Paraformer	**内容审核模型推理**

分类	软件包特性说明	参考文档
算子，包名：AscendCloud-OPP	1. Scatter、Gather算子性能提升，满足MoE训练场景 2. matmul、swiglu、rope等算子性能提升，支持vllm推理场景 3. 支持random随机数算子，优化FFN算子，满足AIGC等场景 4. 支持自定义交叉熵融合算子，满足BMTrain框架训练性能要求 5. 优化PageAttention算子，满足vllm投机推理场景 6. 支持CopyBlocks算子，满足vllm框架beam search解码场景 7. 支持AdvanceStep算子，满足vllm投机推理场景 8. 多个融合算子支持PTA图模式适配，满足AIGC场景 9. 支持两种版本配套算子包（torch2.1.0和python3.9、torch2.3.1和python3.10）	无

2.4 昇腾云服务 6.3.910 版本说明（推荐）

本文档主要介绍昇腾云服务6.3.910版本配套的镜像地址、软件包获取方式和支持的特性能力。

当前版本仅适用于华为公有云。

配套的基础镜像

芯片	镜像地址	获取方式	镜像软件说明	配套关系
Snt 9B	西南-贵阳一 **PyTorch**： swr.cn-southwest-2.myhuaweicloud.com/atelier/pytorch_2_1_ascend:pytorch_2.1.0-cann_8.0.rc3-py_3.9-hce_2.0.2409-aarch64-snt9b-20241112192643-c45ac6b	镜像发布到SWR，从SWR拉取	固件驱动：23.0.6 CANN：cann_8.0.rc3 容器镜像OS：hce_2.0 PyTorch：pytorch_2.1.0、pytorch_2.2.0 MindSpore：MindSpore 2.3.0 FrameworkPTAdapter：6.0.RC3	如果用到CCE，版本要求是CCE Turbo v1.28及以上
300 iDUO	西南-贵阳一 **PyTorch**： swr.cn-southwest-2.myhuaweicloud.com/atelier/pytorch_2_1_ascend:pytorch_2.1.0-cann_8.0.rc3-py_3.9-hce_2.0.2406-aarch64-snt3p-20240906180137-154bd1b	镜像发布到SWR，从SWR拉取	固件驱动:24.1.rc2.3 CANN：cann_8.0.rc3 容器镜像OS：hce_2.0 PyTorch：pytorch_2.1.0 MindSpore lite：2.3.0 FrameworkPTAdapter：6.0.RC3	-

软件包获取地址

软件包名称	软件包说明	获取地址
AscendCloud-6.3.910-xxx.zip	包含 1. 三方大模型训练和推理代码包：AscendCloud-LLM 2. AIGC代码包：AscendCloud-AIGC 3. CV代码包：AscendCloud-CV 4. 算子依赖包：AscendCloud-OPP	获取路径：**Support-E**，在此路径中查找下载ModelArts 6.3.910 版本。 **说明** 如果上述软件获取路径打开后未显示相应的软件信息，说明您没有下载权限，请联系您所在企业的华为方技术支持下载获取。

支持的特性

表 2-4 本版本支持的特性说明

分类	软件包特性说明	参考文档
三方大模型，包名：AscendCloud-LLM	支持如下模型适配PyTorch-NPU的训练(ModelLink) 1. llama2-7b 2. llama2-13b 3. llama2-70b 4. qwen-7b 5. qwen-14b 6. qwen-72b 7. baichuan2-13b 8. chatglm3-6b 9. llama3-8b 10.llama3-70b 11.yi-6B 12.yi-34B 13.qwen1.5-7B 14.qwen1.5-14B 15.qwen1.5-32B 16.qwen1.5-72B 17.qwen2-0.5b 18.qwen2-1.5b 19.qwen2-7b 20.qwen2-72b 21.glm4-9b 22.mistral-7b 23.mixtral-8x7b 24.llama3.1-8b 25.llama3.1-70b 26.qwen2.5-0.5b 27.qwen2.5-7b 28.qwen2.5-14b 29.qwen2.5-32b 30.qwen2.5-72b 31.llama3.2-1b 32.llama3.2-3b 支持如下模型适配PyTorch-NPU的训练(LlamaFactory)	**LLM开源大模型基于DevServer适配ModelLinkPyTorch NPU训练指导** **LLM开源大模型基于DevServer适配LLamaFactory PyTorch NPU训练指导** **LLM开源大模型基于Standard+OBS适配PyTorch NPU训练指导** **LLM开源大模型基于Standard+OBS+SFS适配PyTorch NPU训练指导** **LLM开源大模型基于Lite Cluster适配PyTorch NPU训练指导**

分类	软件包特性说明	参考文档
	1. llama2-7b 2. llama2-13b 3. llama2-70b 4. llama3-8b 5. llama3-70b 6. llama3.1-8b 7. llama3.1-70b 8. qwen1.5-0.5b 9. qwen1.5-1.8b 10.qwen1.5-4b 11.qwen1.5-7b 12.qwen1.5-14b 13.yi-6b 14.yi-34b 15.qwen2-0.5b 16.qwen2-1.5b 17.qwen2-7b 18.qwen2-72b 19.qwen2_vl-2b 20.qwen2_vl-7b 21.falcon-11B 22.glm4-9b 23.qwen2.5-0.5b 24.qwen2.5-7b 25.qwen2.5-14b 26.qwen2.5-32b 27.qwen2.5-72b 28.llama3.2-1b 29.llama3.2-3b	

分类	软件包特性说明	参考文档
	支持如下模型适配PyTorch-NPU的推理。 1. llama-7B 2. llama-13b 3. llama-65b 4. llama2-7b 5. llama2-13b 6. llama2-70b 7. llama3-8b 8. llama3-70b 9. yi-6b 10.yi-9b 11.yi-34b 12.deepseek-llm-7b 13.deepseek-coder-instruct-33b 14.deepseek-llm-67b 15.qwen-7b 16.qwen-14b 17.qwen-72b 18.qwen1.5-0.5b 19.qwen1.5-7b 20.qwen1.5-1.8b 21.qwen1.5-14b 22.qwen1.5-32b 23.qwen1.5-72b 24.qwen1.5-110b 25.qwen2-0.5b 26.qwen2-1.5b 27.qwen2-7b 28.qwen2-72b 29.qwen2.5-0.5b 30.qwen2.5-1.5b 31.qwen2.5-3b 32.qwen2.5-7b 33.qwen2.5-14b 34.qwen2.5-32b 35.qwen2.5-72b 36.baichuan2-7b	**LLM开源大模型基于Lite Server适配PyTorch NPU推理指导** **LLM开源大模型基于Standard适配PyTorch NPU推理指导** **LLM开源大模型基于Lite Cluster适配PyTorch NPU推理指导**

分类	软件包特性说明	参考文档
	37.baichuan2-13b 38.chatglm2-6b 39.chatglm3-6b 40.glm-4-9b 41.gemma-2b 42.gemma-7b 43.mistral-7b 44.mixtral 8*7B 45.falcon2-11b 46.qwen2-57b-a14b 47.llama3.1-8b 48.llama3.1-70b 49.llama-3.1-405B 50.llama-3.2-1B 51.llama-3.2-3B 52.llava-1.5-7b 53.llava-1.5-13b 54.llava-v1.6-7b 55.llava-v1.6-13b 56.llava-v1.6-34b 57.internvl2-26B 58.internvl2-40B 59.MiniCPM-v2.6 60.deepseek-v2-236B 61.deepseek-coder-v2-lite-16B 62.qwen2-vl-7B 63.qwen-vl 64.qwen-vl-chat 65.MiniCPM-v2 Ascend-vllm支持如下推理特性： 1. 支持分离部署 2. 支持多机推理 3. 支持大小模型投机推理及eagle投机推理 4. 支持chunked prefill特性 5. 支持automatic prefix caching 6. 支持multi-lora特性	

分类	软件包特性说明	参考文档
	7. 支持W4A16、W8A16和W8A8量化 8. 升级vLLM 0.6.0	
AIGC，包名：AscendCloud-AIGC	支持如下框架或模型基于DevServer的PyTorch NPU推理： 1. ComfyUI 2. Diffusers 3. Wav2Lip 4. OpenSora1.2 5. OpenSoraPlan1.0 6. MiniCPM-V2.6 7. FLUX.1 8. Hunyuan-Dit 9. Qwen-VL 10.CogVideoX 11.LLama-VID 12.MiniCPM-V2.0 支持如下框架或模型基于DevServer的PyTorch NPU的训练： 1. Qwen-VL 2. Diffusers 3. Koyha_ss 4. Wav2Lip 5. InternVL2 6. OpenSora1.2 7. OpenSoraPlan1.0 8. CogVideoX 9. LLaVA-NeXT 10.LLaVA 11.MiniCPM-V2.0	**Open-Sora 1.2 基于DevServer适配PyTorch NPU训练推理指导** **CogVideoX基于DevServer适配PyTorch NPU训练推理指导** **LLama-VID基于DevServer适配PyTorch NPU推理指导** **InternVL2基于DevServer适配PyTorch NPU训练指导** **MiniCPM-V2.6基于DevServer适配PyTorch NPU训练推理指导** **Qwen-VL基于DevServer适配PyTorch NPU的Finetune训练指导** **LLaVA-Next基于DevServer适配PyTorch NPU训练指导**
CV，包名：AscendCloud-CV	支持如下模型适配MindSpore Lite的推理： 1. Yolov8 2. Bert	**Yolov8基于DevServer适配MindSpore Lite推理指导** **Bert基于DevServer适配MindSpore Lite推理指导**

分类	软件包特性说明	参考文档
算子，包名：AscendCloud-OPP	1. Scatter、Gather算子性能提升，满足MoE训练场景 2. matmul、swiglu、rope等算子性能提升，支持vllm推理场景 3. 支持random随机数算子，优化FFN算子，满足AIGC等场景 4. 支持自定义交叉熵融合算子，满足BMTrain框架训练性能要求 5. 优化PageAttention算子，满足vllm投机推理场景 6. 支持CopyBlocks算子，满足vllm框架beam search解码场景 7. 支持AdvanceStep算子，满足vllm投机推理场景 8. 多个融合算子支持PTA图模式适配，满足AIGC场景	无

2.5 昇腾云服务 6.3.909 版本说明

本文档主要介绍昇腾云服务6.3.909版本配套的镜像地址、软件包获取方式和支持的特性能力。

当前版本仅适用于华为公有云。

配套的基础镜像

芯片	镜像地址	获取方式	镜像软件说明	配套关系
Snt9B	西南-贵阳一 **PyTorch**： swr.cn-southwest-2.myhuaweicloud.com/atelier/pytorch_2_1_ascend:pytorch_2.1.0-cann_8.0.rc3-py_3.9-hce_2.0.2406-aarch64-snt9b-20240910112800-2a95df3 swr.cn-southwest-2.myhuaweicloud.com/atelier/pytorch_2_2_ascend:pytorch_2.2.0-cann_8.0.rc3-py_3.10-hce_2.0.2406-aarch64-snt9b-20240910150953-6faa0ed	镜像发布到SWR，从SWR拉取	固件驱动：23.0.6 CANN：cann_8.0.rc3 容器镜像OS：hce_2.0 PyTorch：pytorch_2.1.0、pytorch_2.2.0 MindSpore：MindSpore 2.3.0 FrameworkPTAdapter：6.0.RC3	如果用到CCE，版本要求是CCE Turbo v1.28及以上
300iDUO	西南-贵阳一 **PyTorch**： swr.cn-southwest-2.myhuaweicloud.com/atelier/pytorch_2_1_ascend:pytorch_2.1.0-cann_8.0.rc3-py_3.9-hce_2.0.2406-aarch64-snt3p-20240906180137-154bd1b	镜像发布到SWR，从SWR拉取	固件驱动:24.1.rc2.3 CANN：cann_8.0.rc3 容器镜像OS：hce_2.0 PyTorch：pytorch_2.1.0 MindSpore lite：2.3.0 FrameworkPTAdapter：6.0.RC3	-

软件包获取地址

软件包名称	软件包说明	获取地址
AscendCloud-6.3.909-xxx.zip	包含 1. 三方大模型训练和推理代码包：AscendCloud-LLM 2. AIGC代码包：AscendCloud-AIGC 3. CV代码包：AscendCloud-CV 4. 算子依赖包：AscendCloud-OPP	获取路径：**Support-E** **说明** 如果上述软件获取路径打开后未显示相应的软件信息，说明您没有下载权限，请联系您所在企业的华为方技术支持下载获取。

支持的特性

表 2-5 本版本支持的特性说明

分类	软件包特性说明	参考文档
三方大模型，包名：AscendCloud-LLM	支持如下模型适配PyTorch-NPU的训练(ModelLink) 1. llama2-7b 2. llama2-13b 3. llama2-70b 4. qwen-7b 5. qwen-14b 6. qwen-72b 7. baichuan2-13b 8. chatglm3-6b 9. llama3-8b 10.llama3-70b 11.yi-6B 12.yi-34B 13.qwen1.5-7B 14.qwen1.5-14B 15.qwen1.5-32B 16.qwen1.5-72B 17.qwen2-0.5b 18.qwen2-1.5b 19.qwen2-7b 20.qwen2-72b 21.glm4-9b 22.mistral-7b 23.mixtral-8x7b 24.llama3.1-8b 25.llama3.1-70b 支持如下模型适配PyTorch-NPU的训练(LlamaFactory) 1. llama2-7b 2. llama2-13b 3. llama2-70b 4. llama3-8b 5. llama3-70b 6. llama3.1-8b 7. llama3.1-70b	**LLM开源大模型基于DevServer适配ModelLinkPyTorch NPU训练指导** **LLM开源大模型基于DevServer适配LLamaFactory PyTorch NPU训练指导** **LLM开源大模型基于Standard+OBS适配PyTorch NPU训练指导** **LLM开源大模型基于Standard+OBS+SFS适配PyTorch NPU训练指导** **LLM开源大模型基于Lite Cluster适配PyTorch NPU训练指导**

分类	软件包特性说明	参考文档
	8. qwen1.5-0.5b 9. qwen1.5-1.8b 10.qwen1.5-4b 11.qwen1.5-7b 12.qwen1.5-14b 13.yi-6b 14.yi-34b 15.qwen2-0.5b 16.qwen2-1.5b 17.qwen2-7b 18.qwen2-72b 19.qwen2_vl-2b 20.qwen2_vl-7b 21.falcon-11B 22.glm4-9b	

分类	软件包特性说明	参考文档
	支持如下模型适配PyTorch-NPU的推理。 1. llama-7B 2. llama-13b 3. llama-65b 4. llama2-7b 5. llama2-13b 6. llama2-70b 7. llama3-8b 8. llama3-70b 9. yi-6b 10.yi-9b 11.yi-34b 12.deepseek-llm-7b 13.deepseek-coder-instruct-33b 14.deepseek-llm-67b 15.qwen-7b 16.qwen-14b 17.qwen-72b 18.qwen1.5-0.5b 19.qwen1.5-7b 20.qwen1.5-1.8b 21.qwen1.5-14b 22.qwen1.5-32b 23.qwen1.5-72b 24.qwen1.5-110b 25.qwen2-0.5b 26.qwen2-1.5b 27.qwen2-7b 28.qwen2-72b 29.baichuan2-7b 30.baichuan2-13b 31.chatglm2-6b 32.chatglm3-6b 33.glm-4-9b 34.gemma-2b 35.gemma-7b 36.mistral-7b	**LLM开源大模型基于Lite Server适配PyTorch NPU推理指导** **LLM开源大模型基于Standard适配PyTorch NPU推理指导** **LLM开源大模型基于Lite Cluster适配PyTorch NPU推理指导**

分类	软件包特性说明	参考文档
	37.mixtral 8*7B 38.falcon2-11b 39.qwen2-57b-a14b 40.llama3.1-8b 41.llama3.1-70b 42.llama-3.1-405B 43.llava-1.5-7b 44.llava-1.5-13b 45.llava-v1.6-7b 46.llava-v1.6-13b 47.llava-v1.6-34b 48.internvl2-26B 49.MiniCPM-v2.6 50.deepseek-v2-236B 51.deepseek-coder-v2-lite-16B Ascend-vllm支持如下推理特性： 1. 支持分离部署 2. 支持多机推理 3. 支持大小模型投机推理及eagle投机推理 4. 支持chunked prefill特性 5. 支持automatic prefix caching 6. 支持multi-lora特性 7. 支持W4A16、W8A16和W8A8量化 8. 升级vLLM 0.6.0	

分类	软件包特性说明	参考文档
AIGC，包名：AscendCloud-AIGC	支持如下框架或模型基于DevServer的PyTorch NPU推理： 1. ComfyUI 2. Diffusers 3. Wav2Lip 4. OpenSora1.2 5. OpenSoraPlan1.0 6. MiniCPM-V2.6 7. FLUX.1 8. Hunyuan-Dit 9. Qwen-VL 支持如下框架或模型基于DevServer的PyTorch NPU的训练： 1. Qwen-VL 2. Diffusers 3. Koyha_ss 4. Wav2Lip 5. InternVL2 6. OpenSora1.2 7. OpenSoraPlan1.0	**FLUX.1基于DevServer适配PyTorch NPU推理指导** **Hunyuan-DiT基于DevServer部署适配PyTorch NPU推理指导** **InternVL2基于DevServer适配PyTorch NPU训练指导** **MiniCPM-V2.6基于DevServer适配PyTorch NPU训练指导** **Qwen-VL基于DevServer适配PyTorch NPU的Finetune训练指导** **Qwen-VL基于DevServer适配PyTorch NPU的推理指导**
CV，包名：AscendCloud-CV	支持如下模型适配MindSpore Lite的推理： 1. Yolov8	**Yolov8基于DevServer适配MindSpore Lite推理指导**
算子，包名：AscendCloud-OPP	1. Scatter、Gather算子性能提升，满足MoE训练场景 2. matmul、swiglu、rope等算子性能提升，支持vllm推理场景 3. 支持random随机数算子，优化FFN算子，满足AIGC等场景 4. 支持自定义交叉熵融合算子，满足BMTrain框架训练性能要求 5. 优化PageAttention算子，满足vllm投机推理场景 6. 支持CopyBlocks算子，满足vllm框架beam search解码场景	无

2.6 昇腾云服务 6.3.908 版本说明

本文档主要介绍昇腾云服务6.3.908版本配套的镜像地址、软件包获取方式和支持的特性能力。

当前版本仅适用于华为公有云。

配套的基础镜像

镜像地址	获取方式	镜像软件说明	配套关系
西南-贵阳一 **PyTorch**： swr.cn-southwest-2.myhuaweicloud.com/atelier/pytorch_2_1_ascend:pytorch_2.1.0-cann_8.0.rc3-py_3.9-hce_2.0.2312-aarch64-snt9b-20240824153350-cebb080 swr.cn-southwest-2.myhuaweicloud.com/atelier/pytorch_2_2_ascend:pytorch_2.2.0-cann_8.0.rc3-py_3.10-hce_2.0.2312-aarch64-snt9b-20240829092203-4ccf328	镜像发布到SWR，从SWR拉取	固件驱动：23.0.6 CANN：cann_8.0.rc3 容器镜像OS：hce_2.0 PyTorch：pytorch_2.1.0、pytorch_2.2.0 MindSpore：MindSpore 2.3.0 FrameworkPTAdapter：6.0.RC3	如果用到CCE，版本要求是CCE Turbo v1.28及以上

软件包获取地址

软件包名称	软件包说明	获取地址
AscendCloud-6.3.908-xxx.zip	包含 1. 三方大模型训练和推理代码包：AscendCloud-LLM 2. AIGC代码包：AscendCloud-AIGC 3. 算子依赖包：AscendCloud-OPP	获取路径：**Support-E** **说明** 如果上述软件获取路径打开后未显示相应的软件信息，说明您没有下载权限，请联系您所在企业的华为方技术支持下载获取。

支持的特性

表 2-6 本版本支持的特性说明

分类	软件包特性说明	参考文档
三方大模型，包名：AscendCloud-LLM	支持如下模型适配PyTorch-NPU的训练(ModelLink) 1. llama2-7b 2. llama2-13b 3. llama2-70b 4. qwen-7b 5. qwen-14b 6. qwen-72b 7. baichuan2-13b 8. chatglm3-6b 9. llama3-8b 10.llama3-70b 11.yi-6B 12.yi-34B 13.qwen1.5-7B 14.qwen1.5-14B 15.qwen1.5-32B 16.qwen1.5-72B 17.qwen2-0.5b 18.qwen2-1.5b 19.qwen2-7b 20.qwen2-72b 21.glm4-9b 22.mistral-7b 23.mixtral-8x7b 支持如下模型适配PyTorch-NPU的训练(LlamaFactory) 1. llama2-7b 2. llama2-13b 3. llama2-70b 4. llama3-8b 5. llama3-70b 6. llama3.1-8b 7. llama3.1-70b 8. qwen1.5-0.5b 9. qwen1.5-1.8b	**LLM开源大模型基于DevServer适配ModelLinkPyTorch NPU训练指导（6.3.908）** **LLM开源大模型基于DevServer适配LLamaFactory PyTorch NPU训练指导（6.3.908）** **LLM开源大模型基于Standard+OBS适配PyTorch NPU训练指导（6.3.908）** **LLM开源大模型基于Standard+OBS+SFS适配PyTorch NPU训练指导（6.3.908）**

分类	软件包特性说明	参考文档
	10.qwen1.5-4b 11.qwen1.5-7b 12.qwen1.5-14b 13.yi-6b 14.yi-34b 15.qwen2-0.5b 16.qwen2-1.5b 17.qwen2-7b 18.qwen2-72b 19.falcon-11B 20.glm4-9b	

分类	软件包特性说明	参考文档
	支持如下模型适配PyTorch-NPU的推理。 1. llama-7B 2. llama-13b 3. llama-65b 4. llama2-7b 5. llama2-13b 6. llama2-70b 7. llama3-8b 8. llama3-70b 9. yi-6b 10.yi-9b 11.yi-34b 12.deepseek-llm-7b 13.deepseek-coder-instruct-33b 14.deepseek-llm-67b 15.qwen-7b 16.qwen-14b 17.qwen-72b 18.qwen1.5-0.5b 19.qwen1.5-7b 20.qwen1.5-1.8b 21.qwen1.5-14b 22.qwen1.5-32b 23.qwen1.5-72b 24.qwen1.5-110b 25.qwen2-0.5b 26.qwen2-1.5b 27.qwen2-7b 28.qwen2-72b 29.baichuan2-7b 30.baichuan2-13b 31.chatglm2-6b 32.chatglm3-6b 33.glm-4-9b 34.gemma-2b 35.gemma-7b 36.mistral-7b	**LLM开源大模型基于DevServer适配PyTorch NPU推理指导（6.3.908）** **LLM开源大模型基于Standard适配PyTorch NPU推理指导（6.3.908）**

分类	软件包特性说明	参考文档
	37.mixtral 8*7B 38.falcon2-11b 39.qwen2-57b-a14b 40.llama3.1-8b 41.llama3.1-70b 42.llama-3.1-405B 43.llava-1.5-7b 44.llava-1.5-13b 45.llava-v1.6-7b 46.llava-v1.6-13b 47.llava-v1.6-34b ascend-vllm支持如下推理特性： 1. 支持分离部署 2. 支持多机推理 3. 支持投机推理 4. 支持chunked prefill特性 5. 支持automatic prefix caching 6. 支持multi-lora特性 7. 支持W4A16、W8A16和W8A8量化	
AIGC，包名：AscendCloud-AIGC	支持如下框架或模型基于DevServer的PyTorch NPU推理： 1. Wav2Lip 2. OpenSora1.2 3. OpenSoraPlan1.0 支持如下框架或模型基于DevServer的PyTorch NPU的训练： 1. Diffusers 2. Kohya_ss 3. Wav2Lip 4. InternVL2 5. OpenSora1.2 6. OpenSoraPlan1.0	**SDXL基于Standard适配PyTorch NPU的LoRA训练指导（6.3.908）** **SD1.5&SDXL Diffusers框架基于DevServer适配PyTorch NPU训练指导（6.3.908）** **SD1.5&SDXL Kohya框架基于DevServer适配PyTorch NPU训练指导（6.3.908）**

分类	软件包特性说明	参考文档
算子，包名：AscendCloud-OPP	1. Scatter、Gather算子性能提升，满足MoE训练场景 2. matmul、swiglu、rope等算子性能提升，支持vllm推理场景 3. 支持random随机数算子，优化FFN算子，满足AIGC等场景 4. 支持自定义交叉熵融合算子，满足BMTrain框架训练性能要求 5. 优化PageAttention算子，满足vllm投机推理场景 6. 支持CopyBlocks算子，满足vllm框架beam search解码场景	无

2.7 昇腾云服务 6.3.907 版本说明

本文档主要介绍昇腾云服务6.3.907版本配套的镜像地址、软件包获取方式和支持的特性能力。

当前版本仅适用于华为公有云。

配套的基础镜像

镜像地址	获取方式	镜像软件说明	配套关系
西南-贵阳一 **PyTorch**： swr.cn-southwest-2.myhuaweicloud.com/atelier/pytorch_2_1_ascend:pytorch_2.1.0-cann_8.0.rc2-py_3.9-hce_2.0.2312-aarch64-snt9b-20240727152329-0f2c29a **MindSpore**： swr.cn-southwest-2.myhuaweicloud.com/atelier/mindspore_2_3_ascend:mindspore_2.3.0-cann_8.0.rc2-py_3.9-hce_2.0.2312-aarch64-snt9b-20240727152329-0f2c29a	镜像发布到SWR，从SWR拉取	固件驱动：23.0.6 CANN：cann_8.0.rc2 容器镜像OS：hce_2.0 PyTorch：pytorch_2.1.0 MindSpore：MindSpore 2.3.0 FrameworkPTAdapter：6.0.RC2	如果用到CCE，版本要求是CCE Turbo v1.25及以上

软件包获取地址

软件包名称	软件包说明	获取地址
AscendCloud-6.3.907-xxx.zip	包含 1. 三方大模型训练和推理代码包：AscendCloud-LLM 2. AIGC代码包：AscendCloud-AIGC 3. 算子依赖包：AscendCloud-OPP	获取路径：**Support-E** **说明** 如果上述软件获取路径打开后未显示相应的软件信息，说明您没有下载权限，请联系您所在企业的华为方技术支持下载获取。

支持的特性

表 2-7 本版本支持的特性说明

分类	软件包特性说明	参考文档
三方大模型，包名：AscendCloud-LLM	支持如下模型适配PyTorch-NPU的训练(ModelLink) 1. llama2-7b 2. llama2-13b 3. llama2-70b 4. qwen-7b 5. qwen-14b 6. qwen-72b 7. baichuan2-13b 8. chatglm3-6b 9. llama3-8b 10.llama3-70b 11.yi-6B 12.yi-34B 13.qwen1.5-7B 14.qwen1.5-14B 15.qwen1.5-32B 16.qwen1.5-72B 17.qwen2-0.5b 18.qwen2-1.5b 19.qwen2-7b 20.qwen2-72b 21.glm4-9b 支持如下模型适配PyTorch-NPU的训练(LlamaFactory) 1. llama3-8b 2. llama3-70b 3. qwen1.5-0.5b 4. qwen1.5-1.8b 5. qwen1.5-4b 6. qwen1.5-7b 7. qwen1.5-14b 8. yi-6b 9. yi-34b 10.qwen2-0.5b 11.qwen2-1.5b	**LLM开源大模型基于DevServer适配ModelLinkPyTorch NPU训练指导（6.3.907）** **LLM开源大模型基于DevServer适配LLamaFactory PyTorch NPU训练指导（6.3.907）** **LLM开源大模型基于Standard+OBS适配PyTorch NPU训练指导（6.3.907）** **LLM开源大模型基于Standard+OBS+SFS适配PyTorch NPU训练指导（6.3.907）**

分类	软件包特性说明	参考文档
	12.qwen2-7b 13.qwen2-7b 14.falcon-11B	

分类	软件包特性说明	参考文档
	支持如下模型适配PyTorch-NPU的推理。 1. llama-7B 2. llama-13b 3. llama-65b 4. llama2-7b 5. llama2-13b 6. llama2-70b 7. llama3-8b 8. llama3-70b 9. yi-6b 10.yi-9b 11.yi-34b 12.deepseek-llm-7b 13.deepseek-coder-instruct-33b 14.deepseek-llm-67b 15.qwen-7b 16.qwen-14b 17.qwen-72b 18.qwen1.5-0.5b 19.qwen1.5-7b 20.qwen1.5-1.8b 21.qwen1.5-14b 22.qwen1.5-32b 23.qwen1.5-72b 24.qwen1.5-110b 25.qwen2-0.5b 26.qwen2-1.5b 27.qwen2-7b 28.qwen2-72b 29.baichuan2-7b 30.baichuan2-13b 31.chatglm2-6b 32.chatglm3-6b 33.glm-4-9b 34.gemma-2b 35.gemma-7b 36.mistral-7b 37.mixtral 8*7B	**LLM开源大模型基于DevServer适配PyTorch NPU推理指导（6.3.907）** **LLM开源大模型基于Standard适配PyTorch NPU推理指导（6.3.907）**

分类	软件包特性说明	参考文档
	38.falcon2-11b 39.qwen2-57b-a14b 40.llama3.1-8b 41.llama3.1-70b ascend-vllm支持如下推理特性： 1. vLLM版本升级至0.5.0	
AIGC，包名：AscendCloud-AIGC	支持如下框架或模型基于DevServer的PyTorch NPU推理： 1. ComfyUI 2. diffusers 3. LLaVA 4. Qwen-VL 5. Wav2Lip 6. OpenSora1.2 7. OpenSoraPlan1.0 支持如下框架或模型基于DevServer的PyTorch NPU的训练： 1. diffusers 2. kohya_ss 3. LLaVA 4. Wav2Lip 5. OpenSora1.2 6. OpenSoraPlan1.0	**SD3 Diffusers框架基于DevServer适配PyTorch NPU推理指导（6.3.907）** **Open-Sora-Plan1.0基于DevServer适配PyTorch NPU训练推理指导（6.3.907）** **Wav2Lip基于DevServer适配PyTorch NPU推理指导** **Wav2Lip基于DevServer适配PyTorch NPU训练指导**
算子，包名：AscendCloud-OPP	1. Scatter、Gather算子性能提升，满足MoE训练场景 2. matmul、swiglu、rope等算子性能提升，支持vllm推理场景 3. 支持random随机数算子，优化FFN算子，满足AIGC等场景 4. 支持自定义交叉熵融合算子，满足BMTrain框架训练性能要求 5. 优化PageAttention算子，满足vllm投机推理场景 6. 支持CopyBlocks算子，满足vllm框架beam search解码场景	无

2.8 昇腾云服务 6.3.906 版本说明

本文档主要介绍昇腾云服务6.3.906版本配套的镜像地址、软件包获取方式和支持的特性能力。

配套的基础镜像

镜像地址	获取方式	配套关系镜像软件说明	配套关系
PyTorch： 西南-贵阳一 swr.cn-southwest-2.myhuaweicloud.com/atelier/pytorch_2_1_ascend:pytorch_2.1.0-cann_8.0.rc2-py_3.9-hce_2.0.2312-aarch64-snt9b-20240606190017-b881580	镜像发布到SWR，从SWR拉取	固件驱动：23.0.5 CANN：cann_8.0.rc2 容器镜像OS：hce_2.0 PyTorch：pytorch_2.1.0 FrameworkPTAdapter：6.0.RC2	如果用到CCE，版本要求是CCE Turbo v1.25及以上

软件包获取地址

软件包名称	软件包说明	获取地址
AscendCloud-6.3.906-xxx.zip	包含 1. 三方大模型训练和推理代码包：AscendCloud-LLM 2. AIGC代码包：AscendCloud-AIGC 3. 算子依赖包：AscendCloud-OPP	获取路径：**Support-E** **说明** 如果上述软件获取路径打开后未显示相应的软件信息，说明您没有下载权限，请联系您所在企业的华为方技术支持下载获取。

支持的特性

表 2-8 本版本支持的特性说明

分类	软件包特性说明	参考文档
三方大模型，包名：AscendCloud-LLM	支持如下模型适配PyTorch-NPU的训练。 1. llama2-7b 2. llama2-13b 3. llama2-70b 4. qwen-7b 5. qwen-14b 6. qwen-72b 7. baichuan2-13b 8. chatglm3-6b 9. llama3-8b 10.llama3-70b 11.yi-6B 12.yi-34B 13.qwen1.5-7B 14.qwen1.5-14B 15.qwen1.5-32B 16.qwen1.5-72B 17.qwen2-0.5b 18.qwen2-1.5b 19.qwen2-7b 20.qwen2-72b 21.glm4-9b	**LLM开源大模型基于DevServer适配PyTorch NPU训练指导（6.3.906）** **LLM开源大模型基于Standard适配PyTorch NPU训练指导（6.3.906）**

分类	软件包特性说明	参考文档
	支持如下模型适配PyTorch-NPU的推理。 1. llama-7B 2. llama-13b 3. llama-65b 4. llama2-7b 5. llama2-13b 6. llama2-70b 7. llama3-8b 8. llama3-70b 9. yi-6b 10.yi-9b 11.yi-34b 12.deepseek-llm-7b 13.deepseek-coder-instruct-33b 14.deepseek-llm-67b 15.qwen-7b 16.qwen-14b 17.qwen-72b 18.qwen1.5-0.5b 19.qwen1.5-7b 20.qwen1.5-1.8b 21.qwen1.5-14b 22.qwen1.5-32b 23.qwen1.5-72b 24.qwen1.5-110b 25.qwen2-0.5b 26.qwen2-1.5b 27.qwen2-7b 28.qwen2-72b 29.baichuan2-7b 30.baichuan2-13b 31.chatglm2-6b 32.chatglm3-6b 33.glm-4-9b 34.gemma-2b 35.gemma-7b 36.mistral-7b 37.mixtral 8*7B	**LLM开源大模型基于DevServer适配PyTorch NPU推理指导（6.3.906）** **LLM开源大模型基于Standard适配PyTorch NPU推理指导（6.3.906）**

分类	软件包特性说明	参考文档
	ascend-vllm支持如下推理特性： 1. vllm版本升级至0.4.2 2. llama、qwen系列模型支持w8a8、w4a16量化 3. 支持prefix caching、投机推理特性	
AIGC，包名：AscendCloud-AIGC	支持如下框架或模型基于DevServer的PyTorch NPU推理： 1. ComfyUI 2. LLaVA 3. Qwen-VL 4. Wav2Lip 支持如下模型基于DevServer的PyTorch NPU的训练： 1. Qwen-VL 2. LLaVA	**SDXL&SD1.5 ComfyUI插件基于DevServer适配PyTorch NPU推理指导（6.3.906）** **LLaVA模型基于DevServer适配PyTorch NPU推理指导（6.3.906）** **Qwen-VL基于DevServer适配PyTorch NPU的推理指导(6.3.906)** **Wav2Lip基于DevServer适配PyTorch NPU推理指导** **LLaVA模型基于DevServer适配PyTorch NPU训练指导（6.3.906）** **Qwen-VL基于DevServer适配PyTorch NPU训练指导(6.3.906)**
算子，包名：AscendCloud-OPP	1. Scatter、Gather算子性能提升，满足MoE训练场景 2. matmul、swiglu、rope等算子性能提升，支持vllm推理场景 3. 新增random随机数算子，优化FFN算子，满足AIGC等场景 4. 新增自定义交叉熵融合算子，满足BMTrain框架训练性能要求 5. 优化PageAttention算子，满足vllm投机推理场景 6. 新增CopyBlocks算子，满足vllm框架beam search解码场景	无

2.9 昇腾云服务 6.3.905 版本说明

本文档主要介绍昇腾云服务6.3.905版本配套的镜像地址、软件包获取方式和支持的特性能力。

配套的基础镜像

镜像地址	获取方式	镜像软件说明	配套关系
PyTorch： 西南-贵阳一 swr.cn-southwest-2.myhuaweicloud.com/atelier/pytorch_2_1_ascend:pytorch_2.1.0-cann_8.0.rc2-py_3.9-hce_2.0.2312-aarch64-snt9b-20240528150158-b521cc0	镜像发布到SWR，从SWR拉取	固件驱动：23.0.5 CANN：cann_8.0.rc2 容器镜像OS：hce_2.0 PyTorch：pytorch_2.1.0 FrameworkPTAdapter：6.0.RC2	如果用到CCE，版本要求是CCE Turbo v1.25及以上

软件包获取地址

软件包名称	软件包说明	获取地址
AscendCloud-3rdLLM-6.3.905-20240611214128.zip	三方大模型训练和推理代码包	获取路径：**Support-E** **说明** 如果上述软件获取路径打开后未显示相应的软件信息，说明您没有下载权限，请联系您所在企业的华为方技术支持下载获取。
AscendCloud-3rdAIGC-6.3.905-20240529154412.zip	AIGC场景训练和推理代码包	
AscendCloud-LLMFramework-6.3.905-20240611151643.zip	大模型推理框架代码包	
AscendCloud-OPP-6.3.905-20240611170314.zip	算子依赖包	

支持的特性

表 2-9 本版本支持的特性说明

分类	软件包特性说明	参考文档
三方大模型，包名：AscendCloud-3rdLLM	支持如下模型适配PyTorch-NPU的训练。 1. llama2-7b 2. llama2-13b 3. llama2-70b 4. qwen-7b 5. qwen-14b 6. qwen-72b 7. baichuan2-13b 8. chatglm3-6b 9. llama3-8b 10.llama3-70b 11.yi-6B 12.yi-34B 13.qwen1.5-7B 14.qwen1.5-14B 15.qwen1.5-32B 16.qwen1.5-72B	**主流开源大模型（PyTorch）基于DevServer训练指导**

分类	软件包特性说明	参考文档
	支持如下模型适配PyTorch-NPU的推理。 1. llama-7B 2. llama-13b 3. llama-65b 4. llama2-7b 5. llama2-13b 6. llama2-70b 7. llama3-8b 8. llama3-70b 9. yi-6b 10.yi-9b 11.yi-34b 12.deepseek-llm-7b 13.deepseek-coder-instruct-33b 14.deepseek-llm-67b 15.qwen-7b 16.qwen-14b 17.qwen-72b 18.qwen1.5-0.5b 19.qwen1.5-7b 20.qwen1.5-1.8b 21.qwen1.5-14b 22.qwen1.5-32b 23.qwen1.5-72b 24.qwen1.5-110b 25.baichuan2-7b 26.baichuan2-13b 27.chatglm2-6b 28.chatglm3-6b 29.gemma-2b 30.gemma-7b 31.mistral-7b 说明： 当前版本不支持推理量化功能（W4A16，W8A8）	主流开源大模型（PyTorch）基于DevServer推理部署

分类	软件包特性说明	参考文档
AIGC，包名：AscendCloud-3rdAIGC	1. SDXL模型： • Fine-tuning微调支持Standard及DevServer模式 • LoRA微调支持DevServer模式 2. Open-Sora1.0训练支持DevServer模式	**SDXL基于Standard适配PyTorch NPU的Finetune高性能训练指导** **SDXL基于DevServer适配PyTorch NPU的Finetune高性能训练指导** **SDXL基于DevServer适配PyTorch NPU的LoRA训练指导** **Open-Sora基于DevServer适配PyTorch NPU训练指导**
大模型推理框架，包名：AscendCloud-LLMFramework	适配vLLM 0.4.2版本（受限发布）： 1. 仅支持部分三方大模型 2. 不支持prefix caching功能 3. 不支持beam search推理场景，不支持n>1推理场景 4. 不支持chunked prefill	无
算子，包名：AscendCloud-OPP	1. Scatter、Gather算子性能提升，满足MoE训练场景 2. matmul、swiglu、rope等算子性能提升，支持vllm推理场景 3. 新增random随机数算子，优化FFN算子，满足AIGC等场景	无

2.10 昇腾云服务 6.3.904 版本说明

昇腾云服务6.3.904版本发布支持的软件包和能力说明如下，软件包获取路径：**Support-E网站**。

发布包	软件包特性说明	配套说明	备注
昇腾云模型代码	**三方大模型**，包名：AscendCloud-3rdLLM PyTorch框架下支持如下模型训练： 1. llama2-7b 2. llama2-13b 3. llama2-70b 4. qwen-7b 5. qwen-14b 6. qwen-72b 7. baichuan2-13b 8. chatglm3-6b PyTorch框架下支持如下模型推理： 1. llama-7B 2. llama-13b 3. llama-65b 4. llama2-7b 5. llama2-13b 6. llama2-70b 7. llama3-8b 8. llama3-70b 9. yi-6b 10.yi-9b 11.yi-34b 12.deepseek-llm-7b 13.deepseek-coder-instruct-33b 14.deepseek-llm-67b 15.qwen-7b 16.qwen-14b 17.qwen-72b 18.qwen1.5-0.5b 19.qwen1.5-7b 20.qwen1.5-1.8b 21.qwen1.5-14b 22.qwen1.5-32b 23.qwen1.5-72b 24.qwen1.5-110b 25.baichuan2-7b 26.baichuan2-13b 27.chatglm2-6b	配套CANN8.0.RC1镜像	训练参考文档： **LLama2系列（PyTorch）基于DevServer训练指导** **Qwen系列（PyTorch）基于DevServer训练指导** **GLM3-6B（PyTorch）基于DevServer训练指导** **Baichuan3-13B（PyTorch）基于DevServer训练指导** 推理参考文档： **主流开源大模型（PyTorch）基于DevServer推理部署**

发布包	软件包特性说明	配套说明	备注
	28.chatglm3-6b 29.gemma-2b 30.gemma-7b 31.mistral-7b		
	AIGC，包名：ascendcloud-aigc 1. Controlnet插件支持NPU推理（适配ComfyUI） 2. Open-Clip模型昇腾适配 3. SD1.5 Finetune高性能训练 4. moondream2推理适配昇腾 5. BERT、YOLO等8个常用模型适配	配套CANN8.0.RC1镜像	参考文档 **SDXL文生图ComfyUI插件基于DevServer适配NPU推理指导** **Open-Clip基于DevServer适配PyTorch NPU训练指导** **SD1.5文生图Finetune高性能训练适配NPU指导** **moondream2基于DevServer适配PyTorch NPU推理指导**
	大模型推理框架，包名：ascendcloud-llmframework 1. VLLM调度层适配ATB、pybind 2. 支持LLAMA7B/13B/65B 3. 支持单机多卡推理 4. ATB模式支持w8a16量化，推理性能提升	配套CANN8.0.RC1镜像	无

发布包	软件包特性说明	配套说明	备注
	算子，包名：AscendCloud-OPP 1. Scatter、Gather算子性能提升，满足MoE场景 2. 昇腾随机数生成算子与GPU保持一致 3. 支持GroupNorm+transpose+BMM融合算子 4. FFN推理算子支持geglu激活函数 5. 支持配套pybind推理的10+算子（matmul 、swiglu、rope等）	配套CANN8.0.RC1镜像	无
基础镜像	CANN8.0.RC1商发版本 **MindSpore**： 西南-贵阳一：swr.cn-southwest-2.myhuaweicloud.com/atelier/mindspore_2_3_ascend:mindspore_2.3.0-cann_8.0.rc1-py_3.9-hce_2.0.2312-aarch64-snt9b-20240516142953-ca51f42 华北-乌兰察布一：swr.cn-north-9.myhuaweicloud.com/atelier/mindspore_2_3_ascend:mindspore_2.3.0-cann_8.0.rc1-py_3.9-hce_2.0.2312-aarch64-snt9b-20240516142953-ca51f42 **PyTorch**： 西南-贵阳一：swr.cn-southwest-2.myhuaweicloud.com/atelier/pytorch_2_1_ascend:pytorch_2.1.0-cann_8.0.rc1-py_3.9-hce_2.0.2312-aarch64-snt9b-20240516142953-ca51f42 华北-乌兰察布一：swr.cn-north-9.myhuaweicloud.com/atelier/pytorch_2_1_ascend:pytorch_2.1.0-cann_8.0.rc1-py_3.9-hce_2.0.2312-aarch64-snt9b-20240516142953-ca51f42	商发镜像发布到SWR	无

3 产品变更公告

3.1 网络调整公告

ModelArts针对网络进行安全加固和优化，新的网络模式可以为用户的资源提供更好的隔离性，提升云上资源的安全。为保障您的网络安全，建议您后续使用新网络创建Standard资源池。

表 3-1 上线局点

上线局点	上线时间
华东二	2024年10月29日 20:00

3.2 预测 API 的域名停用公告

华为云ModelArts将于2024年12月31日 00:00（北京时间）逐步停用预测API的域名huaweicloudapis.com，后续预测API切换使用新域名modelarts-infer.com。

停用范围

影响区域：华为云全部Region

停用影响

新建服务、存量服务停止后再启动、存量服务失败后再启动，会立即切换使用新域名。为保障持续提供推理服务，请您及时更新业务中的预测API的域名。

如果您使用的是VPC内部节点访问ModelArts推理的在线服务，预测API切换域名后，由于内网VPC无法识别公网域名，请**提交工单**联系华为云技术支持打通网络。