ModelArts

服务公告

文档版本	01
发布日期	2025-09-15

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 ModelArts 产品发布公告

本文介绍了AI开发平台ModelArts服务生命周期类产品公告，更多类型的服务公告请参考**服务公告**。

2025 年 6 月

表 1-1 产品公告

序号	公告标题	公告类型	发布时间
1	**华为云华东二、贵阳一局点ModelArts Studio（MaaS）大模型即服务平台-模型体验服务于2025年6月30日-7月30日升级通知**	升级公告	2025年6月28日
2	**华为云AI开发平台ModelArts开发生产（数据准备），资产管理（数据集）于2025年12月31日00:00:00（北京时间）下线通知**	下线公告	2025年6月10日

2025 年 5 月

表 1-2 产品公告

序号	公告标题	公告类型	发布时间
1	**华为云ModelArts旧版镜像计划2025年11月13日0:00（北京时间）下线通知**	下线公告	2025年5月12日
2	**华为云ModelArts系统pip源约束申明公告**	约束申明公告	2025年5月12日

2025 年 4 月

表 1-3 产品下线公告

序号	公告标题	公告类型	发布时间
1	华为云ModelArts旧版推理服务于2025年10月22日00:00（北京时间）启动下线通知	下线公告	2025年4月22日

2024 年

表 1-4 产品下线公告

序号	公告标题	公告类型	发布时间
1	华为云ModelArts自动学习模块于2025年05月23日00:00（北京时间）下线通知	下线公告	2024年11月21日
2	华为云ModelArts自动学习模块的文本分类功能于2024年12月06日00:00（北京时间）下线通知	下线公告	2024年6月5日
3	华为云ModelArts旧版数据集于2024年10月31日00:00（北京时间）下线通知	下线公告	2024年4月30日

2 ModelArts 版本发布说明

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ModelArts7.2.0版本。 ModelArts 7.2.0版本中针对Ascend Snt9b和Snt9b23算力资源的周边依赖组件配套版本关系如下表所示。

ModelArts Lite Server 版本配套关系表

表 2-1 ModelArts Lite Server 版本配套关系表

强依赖组件	Ascend Snt9B配套版本	Ascend Snt9B23配套版本
Lite模式Server节点操作系统	HCE2.0（推荐）/ Ubuntu22.04	HCE2.0（推荐）
NPU固件&驱动	7.5.0.5.220-24.1.0.3（推荐） 7.1.0.9.220-23.0.6	7.5.0.108.220-24.1.rc3.9（推荐） 7.5.0.107.221-24.1.rc3.7
NPU CANN	8.0.1（推荐） 7.0.1.5	8.1.rc2（推荐） 8.0.RC3
CES Agent	2.7.6.6	2.7.6.6

ModelArts Lite Cluster 版本配套关系表

表 2-2 ModelArts Lite Cluster 版本配套关系表

强依赖组件	Ascend Snt9B配套版本	Ascend Snt9B23配套版本
CCE	1.31（推荐）/ 1.28/1.25/1.23（存量）	1.31（推荐）/1.28/1.25/1.23（存量）
Volcano插件	1.18.3	1.18.3

强依赖组件	Ascend Snt9B配套版本	Ascend Snt9B23配套版本
ModelArts Device-Plugin	7.2.2-20250904170414	7.2.2-20250904170414
huawei-npu	2.1.53	2.1.53
os-node-agent	7.2.2-20250812115143	7.2.2-20250812115143
Lite模式Cluster节点操作系统	HCE2.0（推荐）/EulerOS 2.10	HCE2.0
NPU固件&驱动	7.5.0.5.220-24.1.0.3（推荐） 7.1.0.9.220-23.0.6	7.5.0.109.220-24.1.rc3.10（推荐） 7.5.0.108.220-24.1.rc3.9
SFS Turbo Client+	24.12.01（受限功能）	24.12.01（受限功能）
CES Agent	2.8.2.1	2.8.2.1

ModelArts Standard 版本配套关系表

表 2-3 ModelArts Standard 版本配套关系表

强依赖组件	Ascend Snt9B配套版本	Ascend Snt9B23配套版本
CCE	1.31	1.31
Volcano插件	1.18.3	1.18.3
ModelArts Device-Plugin	7.2.2-20250904170414	7.2.2-20250904170414
os-node-agent	7.2.2-20250812115143	7.2.2-20250812115143
Standard模式集群节点操作系统	HCE2.0（推荐）/EulerOS 2.10	HCE2.0
NPU固件&驱动	7.5.0.5.220-24.1.0.3（推荐） 7.1.0.9.220-23.0.6	7.5.0.109.220-24.1.rc3.10（推荐） 7.5.0.108.220-24.1.rc3.9
SFS Turbo Client+	24.12.01（受限功能）	24.12.01（受限功能）
CES Agent	2.8.2.1	2.8.2.1

2.2 ModelArts 产品 HDK 版本策略

ModelArts会定期发布HDK套件（Ascend NPU）新版本，进行特性更新、性能优化和问题修复，以提升用户体验和系统稳定性。为了方便您能够体验最新功能、规避已知漏洞或问题，并保障业务的安全性和可靠性，建议定期升级至最新版本的ModelArts HDK套件。

本文中描述的内容适用于ModelArts Standard、Lite Server和Lite Cluster。

表 2-4 ModelArts 轻量算力节点（Lite Server）HDK 版本生命周期表

卡类型	HDK版本	当前状态	HDK版本在ModelArts商用时间	HDK版本ModelArts EOS（停止服务）时间	备注
Snt9b	24.1.0.6	已商用	2025年6月	2026年6月	与24.1.0.3版本相比，仅更新Driver和MCU
	24.1.0.3	已商用	2025年5月	2026年5月	-
	23.0.6	已商用	2024年6月	2025年6月	-
	23.0.5	EOS	2024年3月	2025年3月	-
	23.0.3	EOS	2024年1月	2025年1月	概率性触发时序问题导致作业拉起失败，已于24年12月提前EOS，请尽快升级至24.1.0.3及以上版本
	23.0.2	EOS	2024年1月	2025年1月	-
Snt9b23	24.1.RC3.6	已商用	2025年6月	2026年6月	-
	24.1.RC3.5	已商用	2025年3月	2026年3月	-
Snt3pr	24.1.0.3	已商用	2025年5月	2026年5月	-
	24.1.RC2.3	已商用	2024年12月	2025年12月	-
Snt3pd	24.1.RC2.3	已商用	2025年6月	2026年6月	-
	23.0.1	已商用	2024年9月	2025年9月	-
Snt9a	23.0.3	已商用	2024年6月	2026年6月	-
	23.0.1	EOS	2023年3月	2024年3月	-

表 2-5 ModelArts 轻量算力集群（Lite Cluster）/标准算力集群（Standard Cluster) HDK 版本生命周期表

CPU架构	实例规格类型	实例规格	支持集群类型&版本	HDK版本	当前驱动状态	HDK版本在ModelArts商用时间	HDK版本ModelArts EOS（停止服务）时间
arm64	Ascend	Snt9b	ModelArts Lite Cluster专属资源池：支持CCE Standard集群（v1.31、v1.28、v1.25、v1.23）、支持CCE Turbo集群（v1.31、v1.28、v1.25、v1.23）； ModelArts Standard专属资源池：支持CCE Standard集群（v1.31、v1.28、v1.23）	24.1.0.3	默认版本	2025年3月	2026年3月
				24.1.rc3.1	A2 CHIJI机型专用版本	2025年6月	2026年6月
				24.1.rc2.3	已商用	2024年9月	2025年9月
				23.0.7	已商用	2024年6月	2025年6月
				23.0.6	已商用	2024年6月	2025年6月
				23.0.5	EOS	2024年3月	2025年3月
				23.0.3	EOS	2024年1月	2025年1月
				23.0.2	EOS	2024年1月	2025年1月
				23.0.1	EOS	2024年1月	2025年1月

CPU架构	实例规格类型	实例规格	支持集群类型&版本	HDK版本	当前驱动状态	HDK版本在ModelArts商用时间	HDK版本ModelArts EOS（停止服务）时间
arm64	Ascend	Snt9b23	ModelArts Lite Cluster专属资源池：支持CCE Standard集群（v1.31、v1.28、v1.25、v1.23）、支持CCE Turbo集群（v1.31、v1.28、v1.25、v1.23）； ModelArts Standard专属资源池：支持CCE Standard集群（v1.31、v1.28、v1.23）	24.1.RC3.10	默认版本	2025年8月	2026年8月
				24.1.RC3.9	已商用	2025年7月	2026年7月
				24.1.RC3.7	已商用	2025年6月	2026年6月
				24.1.RC3.5	已商用	2025年4月	2026年4月
				24.1.RC3.3	已商用	2025年3月	2026年3月

CPU架构	实例规格类型	实例规格	支持集群类型&版本	HDK版本	当前驱动状态	HDK版本在ModelArts商用时间	HDK版本ModelArts EOS（停止服务）时间
x86	Ascend	Snt3PR	ModelArts Lite Cluster专属资源池：支持CCE Standard集群（v1.31、v1.28、v1.25、v1.23）、支持CCE Turbo集群（v1.31、v1.28、v1.25、v1.23）； ModelArts Standard专属资源池：支持CCE Standard集群（v1.31、v1.28、v1.23）	24.1.0.1	默认版本	2025年5月	2026年5月
				24.1.RC2.3	已商用	2024年8月	2025年8月

CPU架构	实例规格类型	实例规格	支持集群类型&版本	HDK版本	当前驱动状态	HDK版本在ModelArts商用时间	HDK版本ModelArts EOS（停止服务）时间
arm64	Ascend	Snt3PD	ModelArts Lite Cluster专属资源池：支持CCE Standard集群（v1.31、v1.28、v1.25、v1.23）、支持CCE Turbo集群（v1.31、v1.28、v1.25、v1.23）； ModelArts Standard专属资源池：支持CCE Standard集群（v1.31、v1.28、v1.23）	24.1.0.1	默认版本	2025年5月	2026年5月
				24.1.RC2.3	已商用	2024年8月	2025年8月
				25.0.RC1.b030	300IDUO高密机型专用版本	2025年5月	2026年5月

CPU架构	实例规格类型	实例规格	支持集群类型&版本	HDK版本	当前驱动状态	HDK版本在ModelArts商用时间	HDK版本ModelArts EOS（停止服务）时间
arm64	Ascend	Snt9	ModelArts Lite Cluster专属资源池：支持CCE Standard集群（v1.31、v1.28、v1.25、v1.23）、支持CCE Turbo集群（v1.31、v1.28、v1.25、v1.23）； ModelArts Standard专属资源池：支持CCE Standard集群（v1.31、v1.28、v1.23）	24.1.rc2.3	默认版本	2024年6月	2025年6月
				23.0.6	EOS	2024年9月	2025年9月
				23.0.3	EOS	2024年6月	2025年6月
				23.0.1	EOS	2023年3月	2024年3月

CPU架构	实例规格类型	实例规格	支持集群类型&版本	HDK版本	当前驱动状态	HDK版本在ModelArts商用时间	HDK版本ModelArts EOS（停止服务）时间
x86	GPU	Ant8 Ant1 Vnt1 Tnt004	ModelArts Lite Cluster专属资源池：支持CCE Standard集群（v1.31、v1.28、v1.25、v1.23）、支持CCE Turbo集群（v1.31、v1.28、v1.25、v1.23）； ModelArts Standard专属资源池：支持CCE Standard集群（v1.31、v1.28、v1.23）	515.65.01	默认版本	2023年12月	2024年12月
				470.182.03	EOS	2022年9月	2023年9月

CPU架构	实例规格类型	实例规格	支持集群类型&版本	HDK版本	当前驱动状态	HDK版本在ModelArts商用时间	HDK版本ModelArts EOS（停止服务）时间
x86	GPU	Lnt002 Hnt02	ModelArts Lite Cluster专属资源池：支持CCE Standard集群（v1.31、v1.28、v1.25、v1.23）、支持CCE Turbo集群（v1.31、v1.28、v1.25、v1.23）； ModelArts Standard专属资源池：支持CCE Standard集群（v1.31、v1.28、v1.23）	535.129.03	默认版本	2024年12月	2025年12月

ModelArts 产品 HDK 版本阶段说明

- **版本商用阶段**：HDK商用版本经过充分验证，稳定可靠。您可以将该版本用于生产环境，享受ModelArts的运维保障。
- **版本EOS（停止服务）阶段**：HDK版本EOS之后，ModelArts将不再支持对该版本的资源创建，同时不提供相应的技术支持，包含新特性更新、漏洞/问题修复、补丁升级以及工单指导、在线排查等客户支持，不再适用于ModelArts服务运维保障。

ModelArts 产品 HDK 版本升级

为了方便您体验新特性、规避已知漏洞或问题，建议您定期升级HDK，使用安全、稳定、可靠的HDK版本。HDK版本EOS之后，您将无法获得相应的技术支持以及ModelArts服务运维保障，请您务必及时升级HDK版本。升级过程中遇到问题，请**提交工单**联系华为云技术支持解决。

ModelArts各个产品形态的HDK版本升级指导请参见以下文档。

- **升级Standard专属资源池驱动**

- 升级Lite Cluster资源池驱动
- 升级Lite Server中的昇腾驱动固件版本

2.3 ModelArts 容器镜像 EOS 公告

ModelArts提供了ARM+Ascend、GPU等规格的统一镜像，包括MindSpore、PyTorch。适用于开发环境，模型训练，服务部署，以提升用户体验和系统稳定性。为了方便您能够体验最新功能、规避已知漏洞或问题，并保障业务的安全性和可靠性，建议定期切换至最新版本的统一镜像。

Ascend 镜像

卡类型	预置镜像版本	当前状态	预置镜像商用时间	预置镜像EOS（停止服务）时间
Snt9b23	mindspore_2.6.0rc1-cann_8.1.rc2-py_3.10-hce_2.0.2503-aarch64-snt9b23	已商用	2025年6月	2026年6月
	pytorch_2.1.0-cann_8.1.rc2-py_3.10-hce_2.0.2503-aarch64-snt9b23	已商用	2025年6月	2026年6月
	pytorch_2.1.0-cann_8.0.rc3-py_3.9-hce_2.0.2409-aarch64-snt9b23	已商用	2025年3月	2026年3月
	pytorch_2.1.0-cann_8.0.rc3-py_3.10-hce_2.0.2412-aarch64-snt9b23	已商用	2025年3月	2026年3月
	mindspore_2.4.10-cann_8.0.rc3-py_3.10-hce_2.0.2412-aarch64-snt9b23	已商用	2025年3月	2026年3月
	pytorch_2.5.1-cann_8.0.rc3-py_3.10-hce_2.0.2503-aarch64-snt9b23	已商用	2025年3月	2026年3月
Snt9b	pytorch_2.1.0-cann_8.1.rc1-py_3.10-euler_2.10.11-aarch64-snt9b	已商用	2025年6月	2026年6月
	mindspore_2.6.0rc1-cann_8.1.rc1-py_3.10-euler_2.10.11-aarch64-snt9b	已商用	2025年6月	2026年6月
	mindspore_2.4.10-cann_8.0.0-py_3.10-euler_2.10.11-aarch64-snt9b	已商用	2025年4月	2026年4月

卡类型	预置镜像版本	当前状态	预置镜像商用时间	预置镜像EOS（停止服务）时间
	pytorch_2.1.0-cann_8.0.0-py_3.10-euler_2.10.11-aarch64-snt9b	已商用	2025年4月	2026年4月
	mindspore_2.4.10-cann_8.0.0-py_3.10-euler_2.10.11-aarch64-snt9	已商用	2025年4月	2026年4月
	mindspore_2.4.0-cann_8.0.rc3-py_3.9-euler_2.10.10-aarch64-snt9b	已商用	2024年12月	2025年12月
	pytorch_2.1.0-cann_8.0.rc3-py_3.9-euler_2.10.10-aarch64-snt9b	已商用	2024年12月	2025年12月
	mindspore_2.3.0-cann_8.0.rc2-py_3.9-euler_2.10.7-aarch64-snt9b	已商用	2024年9月	2025年9月
	pytorch_2.1.0-cann_8.0.rc2-py_3.9-euler_2.10.7-aarch64-snt9b	已商用	2024年9月	2025年9月
	pytorch_1.11.0-cann_8.0.rc2-py_3.9-euler_2.10.7-aarch64-snt9b	已商用	2024年9月	2025年9月
	mindspore_2.3.0-cann_8.0.rc1-py_3.9-euler_2.10.7-aarch64-snt9b	已商用	2024年6月	2025年12月
	pytorch_2.1.0-cann_8.0.rc1-py_3.9-euler_2.10.7-aarch64-snt9b	已商用	2024年6月	2025年12月
	pytorch_1.11.0-cann_8.0.rc1-py_3.9-euler_2.10.7-aarch64-snt9b	已商用	2024年6月	2025年12月
	mindspore_2.2.12-cann_7.0.1.1-py_3.9-euler_2.10.7-aarch64-snt9b	已商用	2024年3月	2025年12月
	pytorch_2.1.0-cann_7.0.1.1-py_3.9-euler_2.10.7-aarch64-snt9b	已商用	2024年3月	2025年12月

卡类型	预置镜像版本	当前状态	预置镜像商用时间	预置镜像EOS（停止服务）时间
	pytorch_1.11.0-cann_7.0.1.1-py_3.9-euler_2.10.7-aarch64-snt9b	已商用	2024年3月	2025年12月
	mindspore_2.2.0-cann_7.0.1-py_3.9-euler_2.10.7-aarch64-snt9b	已商用	2023年12月	2025年12月
	pytorch_2.1.0-cann_7.0.1-py_3.9-euler_2.10.7-aarch64-snt9b	已商用	2023年12月	2025年12月
	pytorch_1.11.0-cann_7.0.1-py_3.9-euler_2.10.7-aarch64-snt9b	已商用	2023年12月	2025年12月
	mindspore_2.2.0-cann_7.0.1-py_3.9-euler_2.10.7-aarch64-snt9b	已商用	2023年12月	2025年12月
	pytorch_2.1.0-cann_7.0.1-py_3.9-euler_2.10.7-aarch64-snt9b	已商用	2023年12月	2025年12月
	pytorch_1.11.0-cann_7.0.1-py_3.9-euler_2.10.7-aarch64-snt9b	已商用	2023年12月	2025年12月
	mindspore_2.1.0-cann_6.3.2-py_3.9-euler_2.10.7-aarch64-snt9b	已商用	2023年9月	2025年9月
	mindspore_2.1.0-cann_6.3.2-py_3.7-euler_2.10.7-aarch64-snt9b	EOS	2023年9月	2024年9月
	pytorch_1.11.0-cann_6.3.2-py_3.7-euler_2.10.7-aarch64-snt9b	EOS	2023年9月	2024年9月
	tensorflow_1.15.0-cann_6.3.2-py_3.7-euler_2.10.7-aarch64-d910b	EOS	2023年9月	2024年9月
Snt9	pytorch_2.1.0-cann_8.0.0-py_3.10-euler_2.10.11-aarch64-snt9	已商用	2025年4月	2026年4月

卡类型	预置镜像版本	当前状态	预置镜像商用时间	预置镜像EOS（停止服务）时间
	mindspore_2.4.10-cann_8.0.0-py_3.10-euler_2.10.11-aarch64-snt9	已商用	2025年4月	2026年4月
	mindspore_2.1.0-cann_6.3.2-py_3.9-euler_2.10.7-aarch64-snt9	已商用	2024年7月	2025年12月
	mindspore_2.1.0-cann_6.3.2-py_3.7-euler_2.8.3-aarch64-snt9	EOS	2023年9月	2024年9月
	pytorch_1.11.0-cann_6.3.2-py_3.7-euler_2.8.3-aarch64-snt9	EOS	2023年9月	2024年9月
	pytorch_1.11.0-cann_6.3.2-py_3.7-euler_2.8.3-aarch64-d910-20230727154652-7d74011	EOS	2023年8月	2024年8月
	mindspore_2.1.0-cann_6.3.2-py_3.7-euler_2.8.3-aarch64-d910-20230727154652-7d74011	EOS	2023年8月	2024年8月
	tensorflow_1.15.0-cann_6.3.2-py_3.7-euler_2.8.3-aarch64-d910-20230727154652-7d74011	EOS	2023年8月	2024年8月
	mindspore_1.7.0-cann_5.1.0-py_3.7-euler_2.8.3-aarch64-d910-20220906095910-6531e92	EOS	2022年10月	2023年10月
	tensorflow_1.15-cann_5.1.0-py_3.7-euler_2.8.3-aarch64-d910-20220906095910-6531e92	EOS	2022年10月	2023年10月
	mindspore_2.1.0-cann_6.3.2-py_3.7-euler_2.10.7-aarch64-d910-20230907104152-28db128	EOS	2022年10月	2023年10月
	mindspore_1.9.0-cann_6.0.0-py_3.7-euler_2.8.3-aarch64-d910-20221116111529	EOS	2022年12月	2023年12月
	mindspore_1.10.0-cann_6.0.1-py_3.7-euler_2.8.3-aarch64-d910-20230303173945-815d627	EOS	2023年4月	2024年4月

卡类型	预置镜像版本	当前状态	预置镜像商用时间	预置镜像EOS（停止服务）时间
	mindspore_1.7.0-cann_5.1.0-py_3.7-euler_2.8.3-aarch64-d910-20220906	EOS	2022年10月	2024年10月
	tensorflow_1.15-cann_5.1.0-py_3.7-euler_2.8.3-aarch64-d910-20220906	EOS	2022年10月	2023年10月
Snt3PR&Snt3PD	pytorch_2.1.0-cann_8.1.rc1-py_3.10-hce_2.0.2503-aarch64-snt3p	已商用	2025年6月	2026年6月
	mindspore_2.6.0rc1-cann_8.1.rc1-py_3.10-hce_2.0.2503-aarch64-snt3p	已商用	2025年6月	2026年6月
	mindspore_2.4.10-cann_8.0.0-py_3.10-euler_2.10.11-aarch64-snt3p	已商用	2025年4月	2026年4月
	pytorch_2.1.0-cann_8.0.0-py_3.10-euler_2.10.11-aarch64-snt3p	已商用	2025年4月	2026年4月
	pytorch_2.1.0-cann_8.0.rc3-py_3.9-hce_2.0.2406-aarch64-snt3p	已商用	2024年9月	2025年9月
	mindspore_2.2.10-cann_7.0.0-py_3.9-hce_2.0.2312-x86_64-snt3p	已商用	2024年5月	2025年5月
	mindspore_2.2.12-cann_7.0.1.1-py_3.9-euler_2.10.7-aarch64-snt3p	已商用	2024年5月	2025年5月
	pytorch_2.1.0-cann_7.0.1.1-py_3.9-euler_2.10.7-aarch64-snt3p	已商用	2024年5月	2025年5月

GPU 镜像

卡类型	预置镜像版本	当前状态	预置镜像商用时间	预置镜像EOS（停止服务）时间
Ant8 Ant1 Vnt1 Hnt02 Lnt002 Tnt004	pytorch_2_1:pytorch_2.1.0-cuda_12.1-py_3.10.6-ubuntu_22.04-x86_64-20250603154214-4e60e43	已商用	2025年6月	2026年6月
	pytorch_2.1.0-cuda_12.1-py_3.9.11-ubuntu_22.04-x86_64-20240313165241-219655b	已商用	2024年4月	2025年4月
	mindquantum_0.9.0-cuda_11.6-py_3.9-ubuntu_20.04-x86_64-20231130162729-973e1a6	已商用	2023年12月	2024年12月
	tensorflow_1.15.5-cuda_11.4-py_3.8-ubuntu_20.04-x86_64-20220926141224-041ba2e	已商用	2022年10月	2023年10月
	tensorflow_2.1.0-cuda_10.1-py_3.7-ubuntu_18.04-x86_64-20221121111529-d65d817	EOS	2022年12月	2023年12月
	pytorch_1.8.0-cuda_10.2-py_3.7-ubuntu_18.04-x86_64-20221118143845-d65d817	EOS	2022年12月	2023年12月
	mindspore_1.7.0-cpu-py_3.7-ubuntu_18.04-x86_64-20221118143809-d65d817	EOS	2022年12月	2023年12月
	mindspore_1.2.0-py_3.7-cuda_10.1-ubuntu_18.04-x86_64-20221118143809-d65d817	EOS	2022年12月	2023年12月
	rlstudio1.0.0-ray1.3.0-cuda10.1-ubuntu18.04	EOS	2022年12月	2023年12月
	pytorch_1.4-cuda_10.1-py37-ubuntu_18.04-x86_64-20221118143845-d65d817	EOS	2022年12月	2023年12月
	mindspore_1.2.0-py_3.7-ubuntu_18.04-x86_64-20221118143809-d65d817	EOS	2022年12月	2023年12月
	tensorflow_1.13-cuda_10.0-py_3.7-ubuntu_18.04-x86_64-20221118143845-d65d817	EOS	2022年12月	2023年12月

卡类型	预置镜像版本	当前状态	预置镜像商用时间	预置镜像EOS（停止服务）时间
	ubuntu_18.04-x86_64-20230404095316-7fcd503	EOS	2023年5月	2024年5月
	or_1.0.0-py_3.7-ubuntu_18.04-x86_64-roma-20231009152946-e7b7e70	EOS	2023年11月	2024年11月
	develop-remote-pyspark_2.4.5-py_3.7-cpu-ubuntu_18.04-x86_64-uid1000-20231009154125-e7b7e70	EOS	2023年11月	2024年11月
	pytorch_1.10.2-cuda_10.2-py_3.7-ubuntu_18.04-x86_64-20221118143845-d65d817	EOS	2022年12月	2023年12月
	cuda_10.2-ubuntu_18.04-x86_64-20230404095316-7fcd503	EOS	2023年5月	2024年5月
	mindspore_1.7.0-cuda_10.1-py_3.7-ubuntu_18.04-x86_64-20221118143809-d65d817	EOS	2022年12月	2023年12月
	mindspore_1.7.0-cpu-py_3.7-ubuntu_18.04-x86_64-20221118143809-d65d817	EOS	2022年12月	2023年12月
	tensorflow_2.1.0-cuda_10.1-py_3.7-ubuntu_18.04-x86_64-20220926144607-041ba2e	EOS	2022年10月	2023年10月
	pytorch_1.4-cuda_10.1-py37-ubuntu_18.04-x86_64-20220926104017-041ba2e	EOS	2022年10月	2023年10月
	mindspore_1.2.0-py_3.7-cuda_10.1-ubuntu_18.04-x86_64-20220926104106-041ba2e	EOS	2022年10月	2023年10月
	tensorflow_1.13-cuda_10.0-py_3.7-ubuntu_18.04-x86_64-20220926104358-041ba2e	EOS	2022年10月	2023年10月
	mindspore_1.7.0-cpu-py_3.7-ubuntu_18.04-x86_64-20220926104017-041ba2e	EOS	2022年10月	2023年10月

卡类型	预置镜像版本	当前状态	预置镜像商用时间	预置镜像EOS（停止服务）时间
	mindspore_1.7.0-cuda_10.1-py_3.7-ubuntu_18.04-x86_64-20220926104017-041ba2e	EOS	2022年10月	2023年10月
	pytorch_1.10.2-cuda_10.2-py_3.7-ubuntu_18.04-x86_64-20221008154718-2b3e39c	EOS	2022年10月	2023年10月
	mindspore_1.2.0-py_3.7-ubuntu_18.04-x86_64-20220926104106-041ba2e	EOS	2022年10月	2023年10月
	pytorch_1.8.0-cuda_10.2-py_3.7-ubuntu_18.04-x86_64-20220926104358-041ba2e	EOS	2022年10月	2023年10月
	pytorch_2.1.0-cuda_12.1-py_3.9.11-ubuntu_22.04-x86_64-20240313165241-219655b	EOS	2024年4月	2025年4月
	cuda_10.2-ubuntu_18.04-x86_64-roma-20220822110152-6f46fd1	EOS	2023年9月	2024年9月
	ubuntu_18.04-x86_64-roma-20220822110152-6f46fd1	EOS	2022年9月	2023年9月
	pytorch_1.8.2-cuda_10.2-py_3.7-ubuntu_18.04-x86_64-20230208092918-2ed5e40	EOS	2023年3月	2024年3月
	develop-remote-pyspark_2.4.5-py_3.7-cpu-ubuntu_18.04-x86_64-uid1000-20221222203856-fcc979e	EOS	2022年12月	2023年12月
	pytorch_1.8.2-cuda_11.1-py_3.7-ubuntu_18.04-x86_64-20220926104358-041ba2e	EOS	2022年10月	2023年12月
	tensorflow_2.6.0-cuda_11.2-py_3.7-ubuntu_18.04-x86_64-20220926144521-041ba2e	EOS	2022年10月	2023年12月
	mindspore_1.2.0-py_3.7-ubuntu_18.04-x86_64-20220926104106-041ba2e	EOS	2022年10月	2023年12月

卡类型	预置镜像版本	当前状态	预置镜像商用时间	预置镜像EOS（停止服务）时间
	notebook2.0-user-defined-gpu-cp37	EOS	2022年10月	2023年12月
	pytorch_1.4-cuda_10.1-py37-ubuntu_18.04-x86_64-20220926104017-041ba2e	EOS	2022年10月	2023年12月
	tensorflow_2.1.0-cuda_10.1-py_3.7-ubuntu_18.04-x86_64-20220926144607-041ba2e	EOS	2022年10月	2023年12月
	notebook2.0-user-defined-cpu-cp37	EOS	2022年10月	2023年12月
	mindspore_1.2.0-py_3.7-cuda_10.1-ubuntu_18.04-x86_64-20220926104106-041ba2e	EOS	2022年10月	2023年12月
	pytorch_1.8.0-cuda_10.2-py_3.7-ubuntu_18.04-x86_64-20220926104358-041ba2e	EOS	2022年10月	2023年12月
	mindspore_1.2.0-py_3.7-ubuntu_18.04-x86_64-20220926104106-041ba2e	EOS	2022年10月	2023年12月
	mindspore_1.2.0-py_3.7-cuda_10.1-ubuntu_18.04-x86_64-20220926104106-041ba2e	EOS	2022年10月	2023年12月
	pytorch_1.4-cuda_10.1-py37-ubuntu_18.04-x86_64-20220926104017-041ba2e	EOS	2022年10月	2023年12月
	notebook2.0-user-defined-gpu-cp37	EOS	2022年10月	2023年12月
	notebook2.0-user-defined-cpu-cp37	EOS	2022年10月	2023年12月
	tensorflow_1.13-cuda_10.0-py_3.7-ubuntu_18.04-x86_64-20220926104358-041ba2e	EOS	2022年10月	2023年12月

卡类型	预置镜像版本	当前状态	预置镜像商用时间	预置镜像EOS（停止服务）时间
	pytorch_1.10.2-cuda_10.2-py_3.7-ubuntu_18.04-x86_64-20221008154718-2b3e39c	EOS	2022年10月	2023年12月
	mindspore_1.7.0-cpu-py_3.7-ubuntu_18.04-x86_64-20220926104017-041ba2e	EOS	2022年10月	2023年12月
	mindspore_1.7.0-cuda_10.1-py_3.7-ubuntu_18.04-x86_64-20220926104017-041ba2e	EOS	2022年10月	2023年12月
	tensorflow_2.1.0-cuda_10.1-py_3.7-ubuntu_18.04-x86_64-20220926144607-041ba2e	EOS	2022年10月	2023年12月
	tensorflow_1.13-cuda_10.0-py_3.7-ubuntu_18.04-x86_64-20220926104358-041ba2e	EOS	2022年10月	2023年12月
	mindspore_1.2.0-py_3.7-cuda_10.1-ubuntu_18.04-x86_64-20220926104106-041ba2e	EOS	2022年10月	2023年12月
	mindspore_1.2.0-py_3.7-ubuntu_18.04-x86_64-20220926104106-041ba2e	EOS	2022年10月	2023年12月
	pytorch_1.8.0-cuda_10.2-py_3.7-ubuntu_18.04-x86_64-20220926104358-041ba2e	EOS	2022年10月	2023年12月
	pytorch_1.4-cuda_10.1-py37-ubuntu_18.04-x86_64-20220926104017-041ba2e	EOS	2022年10月	2023年12月
	develop-remote-pyspark_2.4.5-py_3.7-cpu-ubuntu_18.04-x86_64-uid1000-20231009154125-e7b7e70	EOS	2022年11月	2023年12月
	tensorflow_2.1.0-cuda_10.1-py_3.7-ubuntu_18.04-x86_64-20220926144607-041ba2e	EOS	2022年11月	2023年12月

ModelArts 容器镜像版本阶段说明

- 版本EOS（停止服务）阶段：容器镜像EOS之后，ModelArts将不再支持该镜像的资源创建，同时不提供相应的技术支持，包含新特性更新、漏洞/问题修复、补丁升级以及工单指导、在线排查等客户支持，不再适用于ModelArts服务运维保障。

- 版本商用阶段：容器镜像商用版本经过充分验证，稳定可靠。您可以将该版本用于生产环境，享受ModelArts的运维保障。

ModelArts 容器镜像的版本更新

为了方便您体验新特性、规避已知漏洞或问题，建议您定期更新镜像，使用安全、稳定、可靠的镜像版本。容器镜像EOS之后，您将无法获得相应的技术支持以及ModelArts服务运维保障，请您务必及时更新容器镜像。升级过程中遇到问题，请**提交工单**联系华为云技术支持解决。

2.4 ModelArts 标准算力集群(Standard Cluster)的 Kubernetes 版本策略

ModelArts标准算力集群(Standard Cluster)为用户提供云容器引擎提供高度可扩展的、高性能的企业级Kubernetes集群。由于社区定期发布Kubernetes版本，ModelArts会随之发布相应的集群公测和商用版本。本文将为您介绍ModelArts 标准算力集群(Standard Cluster)的Kubernetes版本策略。

- 版本商用阶段：ModelArts标准算力集群(Standard Cluster)商用版本经过充分验证，稳定可靠。您可以将该版本用于生产环境，享受CCE服务SLA保障。
- 版本EOS（停止服务）阶段：ModelArts标准算力集群(Standard Cluster)版本EOS之后，ModelArts将不再支持对该版本的集群创建，同时不提供相应的技术支持，包含新特性更新、漏洞/问题修复、补丁升级以及工单指导、在线排查等客户支持，不再适用于ModelArts服务SLA保障。ModelArts 标准算力集群(Standard Cluster)版本EOS之后，您将无法获得相应的技术支持以及ModelArts服务运维保障，请您务必及时升级ModelArts 标准算力集群(Standard Cluster)版本。升级过程中遇到问题，请**提交工单**联系华为云技术支持解决。

Kubernetes 版本号	当前状态	ModelArts 标准算力集群(Standard Cluster)版本商用时间	ModelArts 标准算力集群(Standard Cluster)版本EOS（停止服务）时间
v1.31	已商用	2025年5月	2027年5月
v1.28	已商用	2024年6月	2026年6月
v1.25	EOS	2023年6月	2025年6月
v1.23	EOS	2022年6月	2024年6月
v1.21	EOS	2022年4月	2024年4月
v1.19	EOS	2021年3月	2023年9月
v1.17	EOS	2020年7月	2023年1月
v1.15	EOS	2019年12月	2022年9月
v1.13	EOS	2019年6月	2022年3月
v1.11	EOS	2018年10月	2021年3月
v1.9	EOS	2018年3月	2020年12月

ModelArts轻量算力集群（Lite Cluster）集群适配周期：ModelArts轻量算力集群（Lite Cluster）产品的CCE集群为通过华为云CCE服务域进行创建，因此版本维护策略详见**CCE侧的Kubernetes版本策略**。

2.5 ModelArts 轻量算力集群（Lite Cluster）/标准算力集群（Standard Cluster)节点操作系统版本配套关系表

ModelArts为轻量算力集群（Lite Cluster）和标准算力集群（Standard Cluster)提供了预置的节点操作系统，以提升用户体验和系统稳定性。为了方便您能够体验最新功能、规避已知漏洞或问题，并保障业务的安全性和可靠性，建议定期切换至最新版本的操作系统。

- 版本EOS（停止服务）阶段：操作系统版本EOS之后，ModelArts将不再支持对该版本的资源创建，同时不提供相应的技术支持，包含新特性更新、漏洞/问题修复、补丁升级以及工单指导、在线排查等客户支持，不再适用于ModelArts服务运维保障。
- 版本商用阶段：操作系统商用版本经过充分验证，稳定可靠。您可以将该版本用于生产环境，享受ModelArts的运维保障。

如下为当前已经发布的机型与操作系统版本的对应关系，请参考：

表 2-6 已发布机型与操作系统版本的对应关系

CPU架构	实例规格类型	实例规格	支持集群类型&版本	操作系统	当前状态	操作系统在ModelArts商用时间	操作系统在ModelArts EOS（停止服务）时间
arm64	Ascend	Snt9b	ModelArts Lite Cluster专属资源池：支持CCE Standard集群（v1.31、v1.28、v1.25、v1.23）、支持CCE Turbo集群（v1.31、v1.28、v1.25、v1.23）； ModelArts Standard专属资源池：支持CCE Standard集群（v1.31、v1.28、v1.23）	Huawei Cloud EulerOS 2.0(ARM)	已商用	2024年12月23日	2029年03月31日
				EulerOS 2.10(ARM)	已商用	2024年07月10日	2026年12月31日

CPU架构	实例规格类型	实例规格	支持集群类型&版本	操作系统	当前状态	操作系统在ModelArts商用时间	操作系统在ModelArts EOS（停止服务）时间
arm64	Ascend	Snt9b23	ModelArts Lite Cluster专属资源池：支持CCE Standard集群（v1.31、v1.28、v1.25、v1.23）、支持CCE Turbo集群（v1.31、v1.28、v1.25、v1.23）； ModelArts Standard专属资源池：支持CCE Standard集群（v1.31、v1.28、v1.23）	Huawei Cloud EulerOS 2.0(ARM)	已商用	2025年04月09日	2029年03月31日
x86	Ascend	Snt3PR	ModelArts Lite Cluster专属资源池：支持CCE Standard集群（v1.31、v1.28、v1.25、v1.23）、支持CCE Turbo集群（v1.31、v1.28、v1.25、v1.23）； ModelArts Standard专属资源池：支持CCE Standard集群（v1.31、v1.28、v1.23）	Huawei Cloud EulerOS 2.0(x86_64)	已商用	2024年12月23日	2029年03月31日
				EulerOS 2.9(x86_64)	已商用	2023年05月14日	2025年12月30日
					已商用	2024年06月03日	2025年12月30日

CPU架构	实例规格类型	实例规格	支持集群类型&版本	操作系统	当前状态	操作系统在ModelArts商用时间	操作系统在ModelArts EOS（停止服务）时间
arm64	Ascend	Snt3PD	ModelArts Lite Cluster专属资源池：支持CCE Standard集群（v1.31、v1.28、v1.25、v1.23）、支持CCE Turbo集群（v1.31、v1.28、v1.25、v1.23）； ModelArts Standard专属资源池：支持CCE Standard集群（v1.31、v1.28、v1.23）	Huawei Cloud EulerOS 2.0(ARM)	已商用	2025年05月17日	2029年03月31日
				EulerOS 2.9(ARM)	已商用	2024年05月23日	2025年12月30日
x86	GPU	Ant8	ModelArts Lite Cluster专属资源池：支持CCE Standard集群（v1.31、v1.28、v1.25、v1.23）、支持CCE Turbo集群（v1.31、v1.28、v1.25、v1.23）； ModelArts Standard专属资源池：支持CCE Standard集群（v1.31、v1.28、v1.23）	EulerOS 2.10	已商用	2024年01月03日	2026年12月31日

CPU架构	实例规格类型	实例规格	支持集群类型&版本	操作系统	当前状态	操作系统在ModelArts商用时间	操作系统在ModelArts EOS（停止服务）时间
x86	GPU	Ant1	ModelArts Lite Cluster专属资源池：支持CCE Standard集群（v1.31、v1.28、v1.25、v1.23）、支持CCE Turbo集群（v1.31、v1.28、v1.25、v1.23）； ModelArts Standard专属资源池：支持CCE Standard集群（v1.31、v1.28、v1.23）	EulerOS 2.10	已商用	2023年01月20日	2026年12月31日
				EulerOS 2.3	EOS	2021年09月30日	2022年09月30日
					EOS	2021年09月30日	2022年09月30日
x86	GPU	Vnt1 Tnt004 Ant03	ModelArts Lite Cluster专属资源池：支持CCE Standard集群（v1.31、v1.28、v1.25、v1.23）； ModelArts Standard专属资源池：支持CCE Standard集群（v1.31、v1.28、v1.23）	EulerOS 2.9	已商用	2022年07月01日	2025年12月30日

CPU架构	实例规格类型	实例规格	支持集群类型&版本	操作系统	当前状态	操作系统在ModelArts商用时间	操作系统在ModelArts EOS（停止服务）时间
X86	GPU	Hnt02 Lnt002	ModelArts Lite Cluster专属资源池：支持CCE Standard集群（v1.31、v1.28、v1.25、v1.23）、支持CCE Turbo集群（v1.31、v1.28、v1.25、v1.23）； ModelArts Standard专属资源池：支持CCE Standard集群（v1.31、v1.28、v1.23）	Huawei Cloud EulerOS 2.0(x86_64)	已商用	2025年04月09日	2029年03月31日

3 昇腾云版本发布说明

3.1 昇腾云服务 6.5.906 版本说明

本文档主要介绍昇腾云服务6.5.906版本配套的镜像地址、软件包获取方式和支持的特性能力。

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	**pytorch_2.5.1**(适用于大语言模型推理框架和AIGC)： swr.cn-southwest-2.myhuaweicloud.com/atelier/pytorch_ascend:pytorch_2.5.1-cann_8.2.rc1-py_3.11-hce_2.0.2503-aarch64-snt9b-20250729103313-3a25129 **pytorch_2.5.1**(适用于多模态模型推理框架)： swr.cn-southwest-2.myhuaweicloud.com/atelier/pytorch_ascend:pytorch_2.5.1-cann_8.2.rc1-py_3.11-hce_2.0.2503-aarch64-snt9b-20250717151727-8092d23	镜像发布到SWR，从SWR拉取 Region：乌兰一、华东二、西南-贵阳一	固件驱动：24.1.0.6(snt9b)/24.1.rc3.7(snt9b23) CANN：cann_8.2.rc1 容器镜像OS：hce_2.0 PyTorch：pytorch_2.5.1

芯片	镜像地址	获取方式	镜像软件说明
Snt9b23	**pytorch_2.5.1**（适用于大语言模型推理框架和AIGC）： swr.cn-southwest-2.myhuaweicloud.com/atelier/pytorch_ascend:pytorch_2.5.1-cann_8.2.rc1-py_3.11-hce_2.0.2503-aarch64-snt9b23-20250729103313-3a25129		

软件包获取地址

软件包名称	软件包说明	获取地址
AscendCloud-6.5.906-20250820145144.zip	大语言模型推理框架和算子代码包（**Snt9b机型**）	获取路径：**Support-E**，在此路径中查找下载ModelArts 6.5.906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。
AscendCloud-6.5.906-20250820145646.zip	大语言模型推理框架和算子代码包（**Snt9b23机型**）	
AscendCloud-LLMFramework-6.5.906-20250818162611.zip AscendCloud-OPP-6.5.906.A2-20250708143415.zip	多模态模型推理框架和算子代码包（**Snt9b机型**）	

支持的特性

表 3-1 本版本支持的特性说明

分类	软件包特性说明	参考文档
三方大模型推理，包名：AscendCloud-LLM	支持如下模型适配PyTorch-NPU的推理(Ascend-vLLM框架)： 1. DeepSeek-R1-Distill-Qwen-1.5b/7b/8b/14b/32b/70b 2. GLM4-9b 3. qwen2-0.5b/7b/14b/72b/57b-a14b 4. qwen2.5-0.5b/1.5b/3b/7b/14b/32b/72b 5. qwen3-0.6b/1.7b/4b/8b/14b/32b/30b-a3b/235b-a22b 6. QWQ-32b 7. bge-reranker-v2-m3/bge-base-en-v1.5/bge-base-zh-v1.5/bge-large-en-v1.5/bge-large-zh-v1.5/bge-m3 8. qwen2.5VL-7b/32b/72b 9. internvl2.5-26B 10.internvl2-Llama3-76B-AWQ 11.gemma3-27b Ascend-vllm支持如下推理特性： 1. 升级至vLLM 0.9.0 2. 支持多机推理 3. 支持W8A8/AWQ量化 4. 部分模型支持Reasoning Outputs 5. 支持APC 6. 部分模型支持Function Call 7. 支持图模式 说明：具体模型支持的特性请参见大模型推理指导文档	LLM大语言模型推理指导

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. Stable Diffusion 1.5（Diffusers、ComfyUI） 2. Stable Diffusion XL（Diffusers、ComfyUI） 3. Stable Diffusion 3.5（Diffusers、ComfyUI） 4. CogVideoX 5. LLama-VID 6. MiniCPM-V2.0 7. CogVideoX1.5 5b 8. Cogvideo 5b 9. Deepseek Janus-Pro 1b 10.Deepseek Janus-Pro 7b 11.Wan2.1 1.3b 12.Wan2.1 14b 13.自回归模型（VAR/XAR/RandAR/Infinity） 14.Wan2.1-VACE-1.3b 15.HunyuanVideo 支持如下框架或模型基于PyTorch NPU的训练（PyTorch框架） 1. Stable Diffusion 1.5（Diffusers、Kohya_ss） 2. Stable Diffusion XL（Diffusers、Kohya_ss） 3. Wav2Lip 4. InternVL2 5. CogVideoX 6. LLaVA-NeXT 7. LLaVA 8. MiniCPM-V2.0 9. Llama-3.2-11b 10.CogVideoX1.5 5b 11.MiniCPM-V2.6 12.Bunny-Llama-3-8B-V 13.Wan2.1 1.3b 14.Wan2.1 14b	**图像生成模型训练推理** **视频生成模型训练推理** **多模态模型训练推理**

3.2 昇腾云服务 6.5.906.1 版本说明

本文档主要介绍昇腾云服务6.5.906.1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	**pytorch_2.1.0**： swr.cn-southwest-2.myhuaweicloud.com/atelier/pytorch_2_1_ascend:pytorch_2.1.0-cann_8.2.rc1-py_3.11-hce_2.0.2503-aarch64-snt9b-20250729103313-3a25129 **pytorch_2.5.1**： swr.cn-southwest-2.myhuaweicloud.com/atelier/pytorch_ascend:pytorch_2.5.1-cann_8.2.rc1-py_3.11-hce_2.0.2503-aarch64-snt9b-20250729103313-3a25129	镜像发布到SWR，从SWR拉取 Region：乌兰一、华东二、西南-贵阳一	固件驱动：24.1.0.6(snt9b)/24.1.RC3.5(snt9b23) CANN：8.1.RC1.B150 容器镜像OS：hce_2.0 PyTorch：pytorch_2.1.0/pytorch_2.5.1 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
Snt9b23	**pytorch_2.1.0**： swr.cn-southwest-2.myhuaweicloud.com/atelier/pytorch_2_1_ascend:pytorch_2.1.0-cann_8.2.rc1-py_3.11-hce_2.0.2503-aarch64-snt9b23-20250729103313-3a25129 **pytorch_2.5.1**： swr.cn-southwest-2.myhuaweicloud.com/atelier/pytorch_ascend:pytorch_2.5.1-cann_8.2.rc1-py_3.11-hce_2.0.2503-aarch64-snt9b23-20250729103313-3a25129		

软件包获取地址

软件包名称	软件包说明	获取地址
AscendCloud-LLM-6.5.906.1-xxx.zip AscendCloudDriving-6.5.906.1-xxx.zip	包含 1. 三方大模型训练代码包：AscendCloud-LLM 2. 自动驾驶模型代码包：AscendCloudDriving	获取路径：**Support-E**，在此路径中查找下载ModelArts 6.5.906.1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 3-2 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(MindSpeed-LLM) 1. qwen2-0.5b/1.5b/7b/72b 2. glm4-9b 3. llama3.1-8b/70b 4. qwen2.5-0.5b/7b/14b/32b/72b 5. llama3.2-1b/3b 6. qwen3-0.6B/1.7B/4B/8B/14B/32B 7. qwen3-moe-30B-A3B/235B-A22B 8. deepseek-V3/R1 支持如下模型适配PyTorch-NPU的训练(Llama-Factory) 1. qwen2-0.5b/1.5b/7b/72b 2. glm4-9b 3. llama3.1-8b/70b 4. qwen2.5-0.5b/7b/14b/32b/72b 5. llama3.2-1b/3b 6. qwen3-0.6B/1.7B/4B/8B/14B/32B 7. qwen3-moe-30B-A3B/235B-A22B 8. qwen2_vl-2b/7b/72b 9. qwen2.5-vl-7b/72b 10.intern2.5-vl-8b/78b 11.gemma3-27b 支持如下模型适配PyTorch-NPU的训练（VeRL） 1. qwen3-32B 2. qwen2.5-vl-32B	LLM大语言模型训练指导

分类	软件包特性说明	参考文档
自动驾驶，包名：AscendCloudDriving	支持如下模型适配PyTorch-NPU的训练 1. PointPillars 2. MapTRv2 3. sparse4D 4. OpenVLA	自动驾驶模型训练推理

3.3 历史发布版本

3.3.1 昇腾云服务 6.5.905 版本说明

本文档主要介绍昇腾云服务6.5.905版本配套的镜像地址、软件包获取方式和支持的特性能力。

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	swr.cn-southwest-2.myhuaweicloud.com/atelier/pytorch_ascend:pytorch_2.5.1-cann_8.1.rc1-py_3.10-hce_2.0.2503-aarch64-snt9b-20250514161205-a9c5055	镜像发布到SWR，从SWR拉取 Region：西南-贵阳一	固件驱动：24.1.0.6(snt9b)/24.1.RC3.5(snt9b23) CANN：8.1.RC1.B150 容器镜像OS：hce_2.0 PyTorch：pytorch_2.5.1 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
snt9b23	swr.cn-southwest-2.myhuaweicloud.com/atelier/pytorch_ascend:pytorch_2.5.1-cann_8.1.rc1-py_3.10-hce_2.0.2503-aarch64-snt9b23-20250514161205-a9c5055		

软件包获取地址

软件包名称	软件包说明	获取地址
AscendCloud-6.5.905-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 4. 自动驾驶模型代码包：AscendCloud-ACD	获取路径：**Support-E**，在此路径中查找下载ModelArts 6.5.905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 3-3 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(MindSpeed-LLM) 1. qwen2-0.5b/1.5b/7b/72b 2. glm4-9b 3. llama3.1-8b/70b 4. qwen2.5-0.5b/7b/14b/32b/72b 5. llama3.2-1b/3b 6. qwen3-0.6B/1.7B/4B/8B/14B/32B 7. qwen3-moe-30B-A3B/235B-A22B 8. deepseek-V3/R1 支持如下模型适配PyTorch-NPU的训练(Llama-Factory) 1. qwen2-0.5b/1.5b/7b/72b 2. glm4-9b 3. llama3.1-8b/70b 4. qwen2.5-0.5b/7b/14b/32b/72b 5. llama3.2-1b/3b 6. qwen3-0.6B/1.7B/4B/8B/14B/32B 7. qwen3-moe-30B-A3B/235B-A22B 8. qwen2_vl-2b/7b/72b 9. qwen2.5-vl-7b/72b 10.intern2.5-vl-8b/78b 11.gemma3-27b 支持如下模型适配PyTorch-NPU的训练（VeRL） 1. qwen3-32B 2. qwen2.5-vl-32B	LLM大语言模型训练指导

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(Ascend-vLLM框架)： 1. DeepSeek-R1-Distill-Qwen-1.5b/7b/8b/14b/32b/70b 2. GLM4-9b 3. qwen2-0.5b/7b/14b/72b/57b-a14b 4. qwen2.5-0.5b/1.5b/3b/7b/14b/32b/72b 5. qwen3-0.6b/1.7b/4b/8b/14b/32b/30b-a3b/235b-a22b 6. QWQ-32b 7. qwen2.5VL-7b/32b/72b 8. gemma3-27b Ascend-vllm支持如下推理特性： 1. 升级至vLLM 0.8.5 2. 支持多机推理 3. 支持W8A8/AWQ量化 4. 部分模型支持Reasoning Outputs 5. 支持图模式 说明：具体模型支持的特性请参见大模型推理指导文档	**LLM大模型推理指导**

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. Stable Diffusion 1.5（Diffusers、ComfyUI） 2. Stable Diffusion XL（Diffusers、ComfyUI） 3. Stable Diffusion 3（Diffusers） 4. Stable Diffusion 3.5（Diffusers、ComfyUI） 5. Wav2Lip 6. OpenSora1.2 7. OpenSoraPlan1.0 8. Hunyuan-Dit 9. Qwen-VL 10.CogVideoX 11.LLama-VID 12.MiniCPM-V2.0 13.CogVideoX1.5 5b 14.Cogvideo 5b 15.Deepseek Janus-Pro 1b 16.Deepseek Janus-Pro 7b 17.Wan2.1 1.3b 18.Wan2.1 14b 19.自回归模型（VAR/XAR/RandAR/Infinity） 支持如下框架或模型基于PyTorch NPU的训练（PyTorch框架） 1. Qwen-VL 2. Stable Diffusion 1.5（Diffusers、Kohya_ss） 3. Stable Diffusion XL（Diffusers、Kohya_ss）	**图像生成模型训练推理** **视频生成模型训练推理** **多模态模型训练推理** **内容审核模型训练推理**

分类	软件包特性说明	参考文档
	4. Wav2Lip 5. InternVL2 6. OpenSora1.2 7. OpenSoraPlan1.0 8. CogVideoX 9. LLaVA-NeXT 10.LLaVA 11.MiniCPM-V2.0 12.Llama-3.2-11b 13.CogVideoX1.5 5b 14.MiniCPM-V2.6 15.Bunny-Llama-3-8B-V 16.Wan2.1 1.3b 17.Wan2.1 14b	
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 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 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. 算子包支持A3场景，pytorch版本升级至2.5.1	无

分类	软件包特性说明	参考文档
自动驾驶，包名：AscendCloud-ACD	支持如下模型适配PyTorch-NPU的训练： 1. OpenEmma 2. Senna 3. SparseDrive 4. UniAD 5. VAD	**自动驾驶模型训练推理**

3.3.2 昇腾云服务 6.5.902 版本说明

本文档主要介绍昇腾云服务6.5.902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_8.0.0-py_3.10-hce_2.0.2412-aarch64-snt9b-20250207103006-97ebd68 **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	镜像发布到SWR， region：西南-贵阳一， 从SWR拉取	固件驱动：24.1.0.6 CANN：cann_8.0.rc3、8.0.0.B100 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

软件包获取地址

软件包名称	软件包说明	获取地址
AscendCloud-6.5.902-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**，在此路径中查找下载ModelArts 6.5.902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 3-4 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(MindSpeed-LLM) 1. qwen2-0.5b 2. qwen2-1.5b 3. qwen2-7b 4. qwen2-72b 5. glm4-9b 6. llama3.1-8b 7. llama3.1-70b 8. qwen2.5-0.5b 9. qwen2.5-7b 10.qwen2.5-14b 11.qwen2.5-32b 12.qwen2.5-72b 13.llama3.2-1b 14.llama3.2-3b 支持如下模型适配PyTorch-NPU的训练(Llama-Factory) 1. llama3.1-8b 2. llama3.1-70b 3. llama3.2-1b 4. llama3.2-3b 5. qwen2-0.5b 6. qwen2-1.5b 7. qwen2-7b 8. qwen2-72b 9. qwen2.5-0.5b 10.qwen2.5-7b 11.qwen2.5-14b 12.qwen2.5-32b 13.qwen2.5-72b 14.glm4-9b 15.qwen2_vl-2b 16.qwen2_vl-7b 17.qwen2_vl-72b	**LLM开源大模型基于Lite Server适配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
	18.qwen2.5-vl-7b 19.qwen2.5-vl-72b	
	支持如下模型适配PyTorch-NPU的推理(Ascend-vLLM框架)： 1. QwQ-32B 2. DeepSeek-R1-Distill-Llama-8B 3. DeepSeek-R1-Distill-Llama-70B 4. DeepSeek-R1-Distill-Qwen-1.5B 5. DeepSeek-R1-Distill-Qwen-7B 6. DeepSeek-R1-Distill-Qwen-14B 7. DeepSeek-R1-Distill-Qwen-32B 8. bge-reranker-v2-m3 9. internvl2.5-38B 10.qwen2.5-vl-7B 11.qwen2.5-vl-72B 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W4A16、W8A16和W8A8量化 4. 升级vLLM 0.7.2 5. 部分模型支持Reasoning Outputs 说明：具体模型支持的特性请参见大模型推理指导文档	**LLM开源大模型基于Lite Server适配Ascend-vLLM 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
AIGC，包名：AscendCloud-AIGC	支持如下框架或模型基于PyTorch NPU推理（PyTorch框架）： 1. Stable Diffusion 1.5（Diffusers、ComfyUI） 2. Stable Diffusion XL（Diffusers、ComfyUI） 3. Stable Diffusion 3（Diffusers） 4. Stable Diffusion 3.5（Diffusers、ComfyUI） 5. Wav2Lip 6. OpenSora1.2 7. OpenSoraPlan1.0 8. Hunyuan-Dit 9. Qwen-VL 10.CogVideoX 11.LLama-VID 12.MiniCPM-V2.0 13.CogVideoX1.5 5b 14.Cogvideo 5b 15.Deepseek Janus-Pro 1b 16.Deepseek Janus-Pro 7b 17.Wan2.1 1.3b 18.Wan2.1 14b 支持如下框架或模型基于PyTorch NPU的训练（PyTorch框架）： 1. Qwen-VL 2. Stable Diffusion 1.5（Diffusers、Kohya_ss） 3. Stable Diffusion XL（Diffusers、Kohya_ss） 4. Wav2Lip 5. InternVL2 6. OpenSora1.2	**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**

分类	软件包特性说明	参考文档
	7. OpenSoraPlan1.0 8. CogVideoX 9. LLaVA-NeXT 10.LLaVA 11.MiniCPM-V2.0 12.Llama-3.2-11b 13.CogVideoX1.5 5b 14.MiniCPM-V2.6 15.Bunny-Llama-3-8B-V	
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

3.3.3 昇腾云服务 6.5.901 版本说明

本文档主要介绍昇腾云服务6.5.901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
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-6.5.901-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 6.5.901 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 3-5 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(MindSpeed-LLM，原名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Lite Server适配ModelLinkPyTorch NPU训练指导** **LLM开源大模型基于Lite Server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 30.MiniCPM-2B 31.MiniCPM3-4B	

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.internvl2.5-4B 62.internvl2.5-8B 63.internvl2.5-78B 64.MiniCPM-v2.6 65.deepseek-v2-236B 66.deepseek-coder-v2-lite-16B 67.qwen2-vl-2B 68.qwen2-vl-7B 69.qwen2-vl-72B 70.qwen-vl 71.qwen-vl-chat 72.MiniCPM-v2	

分类	软件包特性说明	参考文档
	73.gte-Qwen2-7B-instruct 74.bge-large-en-v1.5 75.bge-base-en-v1.5 76.llava-onevision-qwen2-0.5b-ov-hf 77.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 10.支持 input_embed输入 11.分离部署支持调优工具 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. Hunyuan-Dit 7. Qwen-VL 8. CogVideoX 9. LLama-VID 10.MiniCPM-V2.0 11.SD3 12.SD3.5 13.CogVideoX1.5 5b 14.Cogvideo 5b 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. Kohya_ss 4. Wav2Lip 5. InternVL2 6. OpenSora1.2 7. OpenSoraPlan1.0 8. CogVideoX 9. LLaVA-NeXT 10.LLaVA 11.MiniCPM-V2.0 12.Llama-3.2-11b 13.CogVideoX1.5 5b 14.MiniCPM-V2.6 15.Bunny-Llama-3-8B-V	**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**

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

3.3.4 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
300i DUO	**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 3-6 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.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大语言模型训练文档

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大语言模型推理文档

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. Hunyuan-Dit 7. Qwen-VL 8. CogVideoX 9. LLama-VID 10.MiniCPM-V2.0 11.SD3 12.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. Kohya_ss 4. Wav2Lip 5. InternVL2 6. OpenSora1.2 7. OpenSoraPlan1.0 8. CogVideoX 9. LLaVA-NeXT 10.LLaVA 11.MiniCPM-V2.0 12.Llama-3.2-11b 13.CogVideoX1.5 5b 14.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**
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

3.3.5 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 3-7 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 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. 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（PyTorch框架）： 1. Qwen-VL 2. Diffusers 3. Kohya_ss 4. Wav2Lip 5. InternVL2 6. OpenSora1.2 7. OpenSoraPlan1.0 8. CogVideoX 9. LLaVA-NeXT 10.LLaVA 11.MiniCPM-V2.0 12.Llama-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**
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

3.3.6 昇腾云服务 6.3.910 版本说明

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 3-8 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(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大语言模型训练文档

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大语言模型推理文档

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. Hunyuan-Dit 8. Qwen-VL 9. CogVideoX 10.LLama-VID 11.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. Kohya_ss 4. Wav2Lip 5. InternVL2 6. OpenSora1.2 7. OpenSoraPlan1.0 8. CogVideoX 9. LLaVA-NeXT 10.LLaVA 11.MiniCPM-V2.0	**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**
CV，包名：AscendCloud-CV	支持如下模型适配MindSpore Lite的推理： 1. Yolov8 2. Bert	**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	无

3.3.7 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 3-9 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-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Lite Server适配ModelLinkPyTorch NPU训练指导** **LLM开源大模型基于Lite Server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. Hunyuan-Dit 8. 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. Kohya_ss 4. Wav2Lip 5. InternVL2 6. OpenSora1.2 7. OpenSoraPlan1.0	**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**
CV，包名：AscendCloud-CV	支持如下模型适配MindSpore Lite的推理： 1. Yolov8	**Yolov8基于Lite Server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

4 ModelArts 产品变更公告

4.1 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 4-1 上线局点

上线局点	上线时间
华东二	2024年10月29日 20:00

4.2 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。

5 ModelArts Studio（MaaS）模型发布公告

本文介绍了ModelArts Studio（MaaS）服务的新模型发布记录与特性。

2025 年 07 月

表 5-1 模型发布公告

模型名称	模型类型	模型版本	支持区域	版本说明	预置服务	公共池部署	专属池部署
Kimi-K2	文本对话	V2	西南-贵阳一	新模型上架	X	√	√
Qwen2.5-72B-4K	文本对话	V4	西南-贵阳一	版本更新	X	√	√
BGE-M3	向量模型	V1	西南-贵阳一	新模型上架	√	X	X
DeepSeek-R1-Distill-Llama-70B-32K	文本对话	V4	西南-贵阳一	版本更新	X	√	√
Deepseek-Coder-33B-32K	文本对话	V1	华东二	新模型上架	X	√	√
DeepSeek-V3-64K	文本对话	V1	华东二	新模型上架	X	√	√
Qwen3-32B-64K	文本对话	V1	华东二	新模型上架	X	√	√
Qwen3-235B-A22B-64K	文本对话	V1	华东二	新模型上架	X	√	√
Qwen2.5-VL-72B-32K	图像理解	V5	华东二	版本更新	X	√	√

模型名称	模型类型	模型版本	支持区域	版本说明	预置服务	公共池部署	专属池部署
DeepSeek-R1-64K-0528	文本对话	V1	华北-乌兰察布一	新模型上架	X	√	√