

MaaS 模型即服务

控制台功能

文档版本

01

发布日期

2026-10-12



版权所有 © 华为云计算技术有限公司 2026。保留一切权利。

非经本公司书面许可，任何单位和个人不得擅自摘抄、复制本文档内容的部分或全部，并不得以任何形式传播。

商标声明



HUAWEI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 管理 API Key..... 1
- 2 在模型广场查看预置模型.....7
- 3 使用预置服务..... 15
- 4 使用自定义接入点..... 22
- 5 MaaS 在线体验..... 34
 - 5.1 在 MaaS 体验文本对话..... 34
 - 5.2 在 MaaS 体验图像理解..... 39
 - 5.3 在 MaaS 体验图片生成.....42
 - 5.4 在移动端体验 MaaS 文本对话.....45

1 管理 API Key

在使用MaaS时，用户经常需要通过API接口调用模型服务进行数据请求和模型推理等操作。然而，如果API Key管理不当（如密钥丢失或泄露），可能导致未经授权的访问风险，从而影响服务的安全性和稳定性。如何确保API Key的安全管理，防止未授权访问？为了解决这个问题，MaaS平台提供了API Key管理功能，支持对API Key进行创建、编辑、权限管理等操作，确保只有具备有效API Key的用户才能成功调用模型服务，且只有具备相应权限的用户才能管理API Key，同时减少密钥长期暴露的风险。本文介绍如何对API Key进行创建、编辑和删除等操作。

约束限制

- 数量限制：每个账户最多可同时存在30个有效API Key，超出后需要删除旧API Key才能新建。
- 不可找回：新建API Key后需立即保存，系统不会存储明文，丢失后无法恢复。
- 鉴权时效：删除API Key后，基于该Key的接口调用将立即失效。

前提条件

具有API Key操作权限。关于IAM子用户的API Key权限管理，请参见[管理子用户的API Key操作权限](#)。

创建 API Key

最多可创建30个API Key。API Key仅会在新建后显示一次，请及时复制并妥善保存。如果密钥丢失，请新建API Key。

1. 登录[MaaS控制台](#)，在顶部导航栏中选择目标区域。
2. 在左侧导航栏，选择“管理与统计 > API Key 管理”。
3. 在“API Key管理”页面，单击“创建API Key”，配置基础信息和权限设置，单击“创建”。

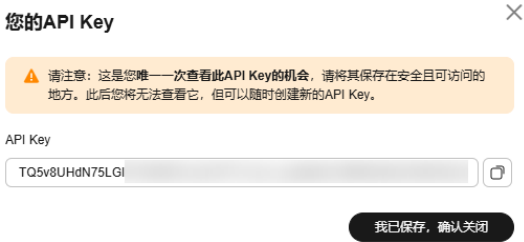
标签在创建完成后，不支持修改。

表 1-1 创建 API Key 参数说明

参数		说明
基础信息	标签	自定义API Key的标签。标签具有唯一性，不可重复。仅支持大小写英文字母、数字、下划线、中划线，长度范围为1~100个字符。
	描述	自定义API Key的描述，长度范围为1~100个字符。
权限设置	权限范围	支持选择“全部（访问所有资源不受限）”或“自定义（指定访问资源范围）”。默认为“全部（访问所有资源不受限）”。
	IP 调用白名单	当“权限范围”选择“自定义（指定访问资源范围）”时，可配置此参数。 开启后，仅白名单中的IP地址可调用模型服务。支持IPv4网段，输入后按Enter生效，支持输入100个地址。IPv4格式示例： • 192.168.1.1 • 192.168.1.10-192.168.1.100 • 192.168.10.0/24 图 1-1 IP 调用白名单 权限设置 权限范围 自定义（指定访问资源范围） ☒ IP 调用白名单 192.168.1.1 × 192.168.1.10-192.168.1.100 × 192.168.10.0/24 × 仅白名单中的IP地址可调用模型服务，您还可以输入 97 个地址。 IP格式示例：192.168.1.1 或 192.168.1.10-192.168.1.100 或 192.168.10.0/24。
	模型服务访问范围	当“权限范围”选择“自定义（指定访问资源范围）”时，可配置此参数。 支持选择“预置服务”或“自定义接入点”。 • 预置服务：使用该API Key通过model参数发起调用时，仅支持访问已勾选的模型及版本。默认“全选”，即选择全部预置服务及版本，包括后续新增预置服务及版本，您也可以按需选择。 • 自定义接入点：默认“全选”，即选择全部自定义接入点，包括后续新增自定义接入点，您也可以按需选择。关于如何创建自定义接入点，请参见使用自定义接入点。

4. 在“您的API Key”对话框，仔细阅读提示信息，复制密钥并保存至安全位置，然后单击“我已保存，确认关闭”。

图 1-2 保存 API Key



关闭后将无法再次查看密钥。创建成功后，在“API Key管理”页面，可以看到新建的API Key，默认显示API Key的前四位和后四位。

图 1-3 API Key



编辑 API Key

API Key创建后，不支持修改标签，其他配置可按需修改。修改权限可能导致当前部分模型服务、自定义接入点等鉴权失败，请确认业务影响后再操作。

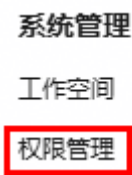
1. 登录[MaaS控制台](#)，在顶部导航栏中选择目标区域。
 2. 在左侧导航栏，选择“管理与统计 > API Key 管理”。
 3. 在目标API Key的“操作”列，单击“编辑”。
 4. 在“编辑API Key”对话框，按需修改基础信息和权限设置，然后单击“确定”。
参数说明请参见[表1-1](#)。
 5. 在“确定编辑API Key”对话框，输入**YES**，单击“确定”。
- API Key修改后，预计五分钟后生效。

管理子用户的 API Key 操作权限

当管理员未配置IAM子用户的API Key操作权限时，MaaS将按照ModelArts控制台的“严格授权模式”开关，判断子用户是否具有API Key的管理权限。

查看或修改“严格授权模式”开关的步骤如下：

1. 管理员登录[ModelArts管理控制台](#)，在旧版控制台左侧导航栏选择“系统管理 > 权限管理”。



说明

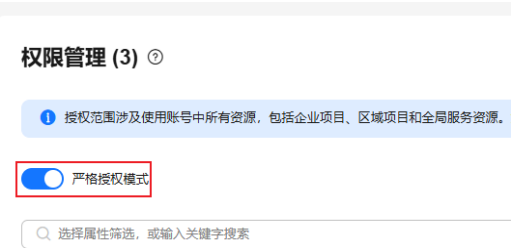
如果您是新版控制台，请在左侧导航栏最下方，单击“返回旧版”。



2. 在“权限管理”页面中按需打开或关闭“严格授权模式”。
- 打开严格授权模式：未配置API Key操作权限的IAM子用户，不具有API Key的管理权限。关于配置IAM子用户的API Key操作权限的步骤，请参见下文。

- 关闭严格授权模式：未配置API Key操作权限的IAM子用户，具有API Key的管理权限。

图 1-4 严格授权模式开关



当管理员通过IAM控制台配置IAM子用户的API Key操作权限时，MaaS将按照管理员的权限配置和ModelArts控制台的“严格授权模式”开关，判断子用户是否具有API Key的管理权限。

配置IAM子用户的API Key操作权限的步骤如下：

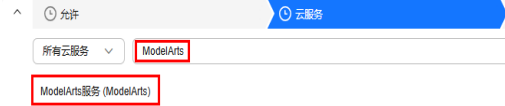
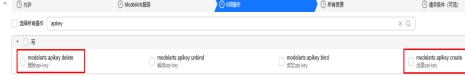
1. 创建自定义策略。
- a. 管理员登录IAM管理控制台，在左侧导航栏选择“权限管理 > 权限”。

b. 在“权限”页面右上角，单击“创建自定义策略”。

c. 在“创建自定义策略”页面，参照下表配置相关信息，单击“确定”。

表 1-2 创建自定义策略参数说明

参数	说明
策略名称	自定义策略名称。只能包含大小写字母、中文、数字、空格或特殊字符(-_.,)。
策略配置方式	单击“可视化视图”。

参数	说明
策略内容	为IAM子用户配置ModelArts服务的API Key权限。 - 选择“允许”或“拒绝”。 - 选择云服务，在搜索框输入ModelArts，选择“ModelArts服务 (ModelArts)”。 图 1-5 云服务配置  - 选择操作，在搜索框输入apikey，按需为IAM子用户选择API Key的操作权限。 “modelarts:apikey:delete”：选中后，IAM子用户可以删除API Key。 “modelarts:apikey:create”：选中后，IAM子用户可以创建或编辑API Key。 图 1-6 操作配置 
策略描述	自定义策略描述。
作用范围	默认为项目级服务。

2. 为用户组添加自定义策略。
- 如果您需要新建用户组，请参见[创建用户组并授权](#)。
- a. 管理员登录IAM管理控制台，在左侧导航栏选择“用户组”。

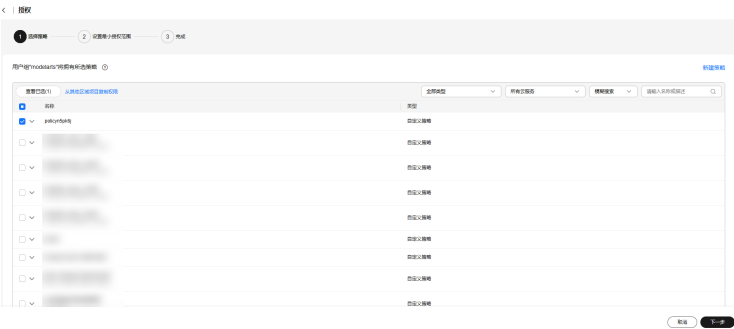
b. 在“用户组”页面，在目标用户组的操作列单击“授权”。

图 1-7 授权用户组



- c. 在“授权”页面，选中1创建的策略名称，单击“下一步”，按需选择授权范围方案，单击“确定”。

图 1-8 授权页面



3. 将IAM子用户加入用户组。
如果您需要新建IAM子用户，请参见[创建IAM用户并登录](#)。
- a. 管理员登录[IAM管理控制台](#)，在左侧导航栏选择“用户组”。

b. 在“用户组”页面，在2修改的用户组的操作列，单击“用户组管理”。

c. 在“用户组管理”对话框，选择目标IAM子用户，单击“确定”。
- 为IAM子用户授权后，IAM子用户在MaaS控制台的“API Key管理”页面，可以使用1选择的操作权限。

删除 API Key

当API Key数量达到上限，或者API Key发生泄露、遗忘等情况时，建议您及时删除不再使用的API Key。**删除后该API Key将无法使用且无法找回，请谨慎操作。**

1. 登录[MaaS控制台](#)，在顶部导航栏中选择目标区域。
2. 在左侧导航栏，选择“管理与统计 > API Key 管理”。
3. 在“API Key管理”页面右侧，单击API Key右侧的“删除”。
4. 在“删除API Key”对话框，单击“确定”。

常见问题

1. 创建API Key后需要等待多久才能生效？
API Key在创建后不会立即生效，通常需要等待几分钟才能生效。
2. API Key是否支持跨区域使用？
API Key是区域级别的，不支持跨区域使用。例如，贵阳一区域的API Key必须通过贵阳一控制台创建，且仅能在该区域内调用服务。其他区域的API Key同理。

2 在模型广场查看预置模型

MaaS平台提供了丰富的开源大模型，在“模型广场”页面可以查看。模型详情页可以查看模型的详细介绍，根据这些信息选择合适的模型进行训练、推理，接入到企业解决方案中。

前提条件

- 已注册华为云账号，详情请见[注册华为云账号](#)。
- 目前三方服务仅支持个人用户开通。

计费说明

调用预置服务：不同类型的模型计费不同，计费详情请参见[MaaS预置模型服务计费项](#)。

模型广场介绍

在模型广场，您可以查看MaaS首推的预置模型类型及模型信息，也可以查看所有的预置模型及相关信息。不同区域支持的模型可能不同，请以控制台实际界面为准。

- “精选模型”页签，查看MaaS首推的预置模型及相关介绍，详情请参考[查看预置模型详细信息](#)。
- “全部模型”页签：查看MaaS支持的所有预置模型及相关介绍，详情请参考[查看预置模型详细信息](#)。

图 2-1 精选模型



图 2-2 全部模型



模型介绍

下表列举了MaaS平台支持的模型清单。关于模型的详细信息请在“模型详情”页面查看。

表 2-1 模型广场的模型系列介绍

模型系列		模型类型	应用场景	支持语言	模型介绍
openPangu	openPangu-2.0-Pro	文本生成	文本生成	中文、英文	openPangu-2.0-Pro 是基于昇腾 NPU 训练的大规模混合专家（MoE）语言模型，参数规模约 505B，每 token 激活参数规模约 18B，模型支持 512k 上下文长度，训练数据总量约34T tokens。后训练阶段完成快慢合一微调（SFT）、多专项强化学习（RL），并通过在线蒸馏（OPD）完成能力合一。
	openPangu-2.0-Flash	文本生成	文本生成	中文、英文	openPangu-2.0-Flash 是基于昇腾 NPU 训练的大规模混合专家（MoE）语言模型，参数规模约 92B，每 token 激活参数规模约 6B，模型支持 512k 上下文长度，训练数据总量约34T tokens。
GLM	GLM-5.2-ArkTS-SPARK	文本生成	推理、编程、智能体任务	中文、英文	GLM-5.2-ArkTS-SPARK是基于 GLM 5.2针对鸿蒙ArkTS的增训模型，更理解鸿蒙ArkTS语法和鸿蒙规范，编写鸿蒙代码语法错误更少，一次性编译率更高，且更加节省Tokens。
	GLM-5.3	文本生成	推理、编程、智能体任务	中文、英文	GLM-5.3 是智谱AI于2026年8月14日发布的旗舰基座模型（7430亿参数，MoE架构，1M上下文窗口）。它与 GLM-5.2 共用同一基座，所有性能提升均来自后训练 Scaling。
	GLM-5.2	文本生成	推理、编程、智能体任务	中文、英文	GLM-5.2是智谱AI 2026年6月17日开源的新一代旗舰大模型，凭借开源SOTA的编码与长程任务能力，显著提升真实开发中的工程稳定性和多端交付可靠性。

模型系列		模型类型	应用场景	支持语言	模型介绍
	GLM-5.1	文本生成	推理、编程、智能体任务	中文、英文	GLM-5.1 是智谱最新旗舰模型，代码能力大大增强，长程任务显著提升，能够在单次任务中持续、自主地工作长达 8 小时，完成从规划、执行到迭代优化的完整闭环，交付工程级成果。 在综合能力与 Coding 能力上，GLM-5.1 整体表现对齐 Claude Opus 4.6，并在长程自主执行、复杂工程优化与真实开发场景中展现出更强的持续工作能力，是构建 Autonomous Agent 与长程 Coding Agent 的理想基座。
Kimi	Kimi-K2.6	文本生成	推理、编程、智能体任务	中文、英文	Kimi K2.6是Kimi最新最强模型，在通用Agent、代码能力上全面提升，并在多个基准测试中取得行业领先成绩。
Deep Seek	DeepSeek-V4.1-Flash	文本生成	对话问答、文本生成推理、图像理解	中文、英文	DeepSeek-V4.1-Flash是一款混合专家（MoE）模型，主干参数量达 5520 亿，并支持高达一百万 token 的上下文长度。
	DeepSeek-V4-Flash	文本生成	对话问答、文本生成推理	中文、英文	DeepSeek-V4-Flash是DeepSeek-V4系列的轻量高效版本，在保持1M超长上下文能力的同时，通过更小的模型参数与激活规模，提供更为快捷、经济的API服务。
通义千问	Qwen 2.5-VL	图像理解	图像理解、对话问答	中文、英文	通义千问2.5-VL是阿里云通义千问团队开源的多模态视觉语言模型，具备强大的视觉和语言理解能力。
	Qwen 3	文本生成	对话问答	中文、英文	Qwen3是Qwen团队研发的大语言模型和大型多模态模型系列，在大规模语言和多模态数据上进行预训练，通过高质量的数据进行后期微调。

模型系列		模型类型	应用场景	支持语言	模型介绍
通义千问Image	Qwen-Image	图像生成	文生图	中文、英文	Qwen-Image的图像生成与编辑通用能力强劲，在文本渲染的场景下表现出色。
	Qwen-Image-Edit	图像生成	文生图、图像编辑	中文、英文	该模型是Qwen-Image的图像编辑版本，合入了其文本渲染能力，支持精准的图中文字修改。
通义万相2.2	Wan2.2-T2V	视频生成	文字生成视频	中文、英文	该文生视频模型采用混合专家（MoE）架构，视频合成更加稳定，支持了更多样化的风格场景。
	Wan2.2-I2V	视频生成	图片生成视频	中文、英文	Wan2.2内置丰富的美学数据集，可轻松定制个性化的电影级画面。
BGE	bge-m3	文本向量化	文本向量化	中文、英文	BGE-M3以其在多语言、多功能和多粒度方面的灵活性而著称。它为超过100种工作语言的语义检索提供了统一的支持，可以同时完成三种常见的检索功能：密集检索、多向量检索和稀疏检索。此外，它还能够处理不同粒度的输入，从短句子到长达8192个Token的长文档。
	bge-reranker-v2-m3	重排序	检索结果再排序	中文、英文	一个轻量级的交叉编码器模型，基于BGE-M3模型开发，具有强大的多语言能力，易于部署，具有快速的推理能力。
爱诗科技	PixVerse V6 T2V	视频生成	文本生成视频	中文、英文	云商店通用商品，由爱诗科技提供的视频大模型API服务。文字生成视频，支持15秒长视频、1080P及多语言输入。

模型系列		模型类型	应用场景	支持语言	模型介绍
	PixVerse V6 IT2V	视频生成	文本+图片生成视频	中文、英文	云商店通用商品，由爱诗科技提供的视频大模型API服务。文字+图像生成视频，支持15秒长视频、1080P及多语言输入。
	PixVerse V6 KF2V	视频生成	文本+图片生成视频	中文、英文	云商店通用商品，由爱诗科技提供的视频大模型API服务。首尾帧生成视频，支持15秒长视频、1080P及多语言输入。
	PixVerse V6 R2V	视频生成	文本+图片生成视频	中文、英文	云商店通用商品，由爱诗科技提供的视频大模型API服务。参考生成视频，支持15秒长视频、1080P及多语言输入。
生数科技	ViduQ 3-Pro T2V	视频生成	文本生成视频	中文	云商店通用商品，由生数科技提供的视频大模型API服务。文字生成视频，支持16秒长视频、1080P及24 FPS。
	ViduQ 3-Pro IT2V	视频生成	文本+图片生成视频	中文	云商店通用商品，由生数科技提供的视频大模型API服务。文字+图像生成视频，支持16秒长视频、1080P及24 FPS。
	ViduQ 3-Pro KF2V	视频生成	文本+图片生成视频	中文	云商店通用商品，由生数科技提供的视频大模型API服务。首尾帧生成视频，支持16秒长视频、1080P及24 FPS。
	ViduQ 3-Turbo T2V	视频生成	文本生成视频	中文	云商店通用商品，由生数科技提供的视频大模型API服务。文字生成视频，支持16秒长视频、1080P及24 FPS。
	ViduQ 3-Turbo IT2V	视频生成	文本+图片生成视频	中文	云商店通用商品，由生数科技提供的视频大模型API服务。文字+图像生成视频，支持16秒长视频、1080P及24 FPS。
	ViduQ 3-Turbo KF2V	视频生成	文本+图片生成视频	中文	云商店通用商品，由生数科技提供的视频大模型API服务。首尾帧生成视频，支持16秒长视频、1080P及24 FPS。

模型系列		模型类型	应用场景	支持语言	模型介绍
	ViduQ 3-Turbo R2V	视频生成	文本+图片生成视频	中文	云商店通用商品，由生数科技提供的视频大模型API服务。参考生成视频，支持16秒长视频、1080P及24 FPS。
	ViduQ 3-Mix R2V	视频生成	文本+图片生成视频	中文	云商店通用商品，由生数科技提供的视频大模型API服务。参考生成视频，支持15秒长视频、1080P及24 FPS。
	ViduQ 3-AD R2V	视频生成	文本+图片生成视频	中文	云商店通用商品，由生数科技提供的视频大模型API服务。参考生成视频，支持15秒长视频、1080P及24 FPS。

查看预置模型详细信息

1. 登录[MaaS控制台](#)，在顶部导航栏中选择目标区域。

2. 在左侧导航栏，单击“模型广场”。

3. 在模型广场右上角，单击“精选模型”或“全部模型”。
- 在“全部模型”页签左侧的“筛选”区域，您可以按需选择模型类型、上下文长度、高级能力、模型系列进行筛选，或者直接在搜索框输入模型名称进行搜索。
4. 单击模型卡片，在“模型详情”页面查看模型的详细信息，例如模型介绍、版本信息、支持的能力等。在“模型详情”页面右上角，可以看到模型支持的操作，操作说明如下。

- [模型对比](#)

- [在线体验](#)

- [API调用](#)

 说明

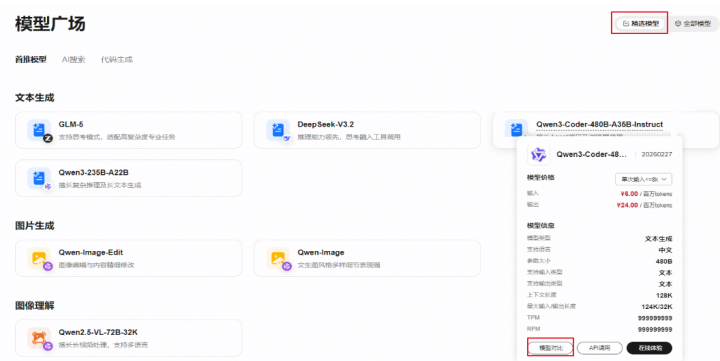
不同的模型支持的操作可能不同，请以控制台显示为准。

模型对比

- MaaS支持对比不同模型或者同一模型的不同版本，可同时对比三个模型/版本。
- 任选以下方式进入模型对比，查看三个模型/版本的模型能力、模型价格、模型限制等区别。相关操作说明参见[表2-2](#)。
- “精选模型” 页签：

- 指定模型与其它模型对比：在首推模型区域下方，鼠标悬浮于模型卡片，单击“模型对比”。

图 2-3 精选模型-模型对比



- 热门模型对比：在首推模型区域右上方，单击“热门模型对比”。
- “全部模型” 页签
 - 指定模型与其它模型对比：单击模型卡片，在“模型详情”页面右上角，单击“模型对比”。
 - 热门模型对比：在首推模型区域右上方，单击“热门模型对比”。

表 2-2 模型对比操作说明

操作	说明
切换模型/版本	MaaS会自动选择模型进行对比，您可以在模型名称右侧单击图标，按需选择模型和版本，单击“确定”。
查看详情	在模型名称下方，单击“查看详情”，会跳转至“模型详情”页面，可以查看该模型的介绍、版本及能力等信息。
API调用	在模型名称下方，单击“API调用”，按照“API调用说明”面板开通或调用模型服务。 开通模型服务后才可体验/调用模型。开通服务将同步开通该服务下所有版本。 关于API Key参数说明，请参见 创建API Key ；关于API调用的参数说明，请参见 模型API调用规范 。
在线体验	在模型名称下方，单击“在线体验”，将跳转至对应的模型体验页面。更多信息，请参见 MaaS在线体验 。 当对比模型为2~3个支持在线体验的文本生成模型时，“模型对比”页面右上角会出现“模型体验对比”，单击后跳转至“文本对话”页面进行模型在线体验对比。如果未开通预置模型服务，会弹出对话框，请按照对话框提示进行开通。
退出对比	在“模型对比”页面右上角，单击“退出对比”。

在线体验

开通模型服务后才可体验/调用模型。开通服务将同步开通该服务下所有版本。开通后，可前往“在线推理 > 预置服务”管理所有模型服务，也可以查看服务的API调用、调用统计等。

任选以下方式进入在线体验。关于在线体验的详细说明，请参见[MaaS在线体验](#)。

- “精选模型” 页签：
 - 模型卡片有操作按钮：单击“在线体验”。
 - 模型卡片无操作按钮，鼠标悬浮于模型卡片，单击“在线体验”。
- “全部模型” 页签：在模型卡片单击“在线体验”。
- “模型详情” 页面：在页面右上角单击“在线体验”，选择版本，或者在版本区域右侧单击“在线体验”。

如果您首次使用，会弹出“开通模型服务”对话框，请仔细查看相关信息，勾选“我已阅读并同意上述说明，及《MaaS 模型即服务声明》”，单击“确认开通”，跳转至模型对应的体验页面进行在线体验。

API 调用

MaaS提供商用级别的推理API服务，无需等待部署，开通模型服务后您可直接体验或调用模型服务，按用量计费（按Token/图片生成时长等）。

1. 任选以下方式进入API调用。
 - “精选模型” 页签：
 - 模型卡片有操作按钮：单击“API调用”。
 - 模型卡片无操作按钮：鼠标悬浮于模型卡片，单击“API调用”。
 - “全部模型” 页签：在模型卡片单击“API调用”。
 - “模型详情” 页面：在页面右上角单击“API调用”，选择版本，或者在版本区域右侧单击“API调用”。
2. 在“API调用说明”面板，按照“API调用说明”面板开通或调用模型服务。

开通模型服务：在“开通模型服务”区域，仔细查看相关信息，勾选“我已阅读并同意上述说明，及《MaaS 模型即服务声明》”，单击“一键开通”。

关于API Key参数说明，请参见[创建API Key](#)；关于API调用的参数说明，请参见[模型API调用规范](#)。

图 2-4 API 调用说明



3 使用预置服务

MaaS预置服务为企业用户提供高性能、高可用的推理API服务，支持按Token用量计费的模式。该服务适用于需要商用级稳定性、更高调用频次和专业支持的场景。

- 企业智能客服：企业希望利用推理API优化客服系统，实现智能问答、意图识别，提升客服效率与客户满意度。
- 内容创作辅助：媒体、广告公司借助理AI进行文案创作、创意生成，提高内容产出的效率与质量。
- 智能数据分析：金融、电商企业通过推理API对海量数据深度分析，挖掘数据价值，辅助决策制定。

约束限制

- 该功能仅支持“西南-贵阳一”区域。
- 开通预置服务时，将自动开通该服务下所有版本，不支持单独开通某版本。
- 暂不支持关闭预置服务。
- 目前三方服务仅支持个人用户开通。

计费说明

在调用模型推理服务的过程中，输入内容首先会被分词（tokenize），转换为模型可识别的Token。在调用MaaS预置服务时，将根据实际使用的Tokens数量进行计费。详细信息，请参见[推理服务计费项](#)。

优惠券说明

- 当有优惠折扣时，预置服务页签会出现相关提示。开通预置服务时，会默认领取可用的优惠券。在扣费时，会优先抵扣优惠券。
- 不同优惠券活动的适用范围和领取条件各不相同，能否成功领取请以实际活动规则为准。
- 模型服务的优惠折扣的发放和使用情况，请前往“费用中心 > 优惠折扣 > 优惠券”进行查看。

服务 API 调用

请求可能会根据实际情况路由到其他区域实例。

开通预置服务

1. 登录[MaaS控制台](#)，在顶部导航栏中选择目标区域。
2. 在左侧导航栏，选择“模型推理 > 在线推理”。
3. 在“预置服务”页签，在目标服务右侧的“操作”列，单击“开通服务”。
4. 在开通预置模型服务对话框，按需勾选预置服务。请仔细查看相关声明或条款，同意相关声明或条款后，单击“一键开通”，预置服务列表页中的“状态”变为“开通”。

勾选目标预置服务后，会自动开通该服务下的所有模型版本。

已开通的预置服务示例如下。预置服务列表的参数说明请参见下表。

图 3-1 已开通预置服务

服务名称来源ID	状态	类型	计费方式	资源定价	网络流量包	创建时间	模型规格	操作
GLM-4.1-7B-Chat-16K-Preview... 70078869c70-4d58-455c...	●	文本生成	按量计费	输入 ¥100~40.00 百万tokens 输出 ¥24.00~420.00 百万tokens 提示词中嵌入 ¥1.30~¥2.00 百	-	-	TPM 800,000 PPM 1,500	查看规格 升级 更多

表 3-1 预置服务列表参数说明

参数	说明
服务名称/资源ID	预置服务的名称和ID。 - 查看服务的版本和model参数：在服务名称左侧单击  图标，可以查看该服务的版本。“model参数”列显示的名称可用于模型调用时使用，即model参数的值。 图 3-2 查看服务版本  单击版本名称，将跳转至“模型详情”页面，您可以查看模型版本的详细信息。 - 查看服务的账单：在预置服务ID后，单击  图标复制ID，在 费用中心 的“账单管理 > 流水和明细账单”页面的“明细账单”页签，通过资源ID查看账单详情。
状态	- 开通：已开通预置服务。 - 未开通：未开通预置服务。
类型	预置服务的类型。

图 3-2 查看服务版本



参数	说明
计费方式	预置服务的计费方式，不同模型的计费方式可能不同，请以实际环境为准。更多信息，请参见 推理服务计费项 。
推理定价	预置服务的推理定价，不同模型的推理定价可能不同，请以实际环境为准。更多信息，请参见推理服务计费项。 在页面右上角，您可以按需打开/关闭“百万tokens单位展示”。 图 3-3 百万 tokens 单位展示  - 关闭“百万tokens单位展示”：将按照千tokens为单位展示定价。 图 3-4 千 tokens 单位示例  - 打开“百万tokens单位展示”：将按照百万tokens为单位展示定价。 图 3-5 百万 tokens 单位示例 

参数	说明
优惠折扣	- 模型服务的优惠折扣的发放和使用情况，请前往费用中心查看。 - 预置服务已有的优惠折扣，“--”表示没有优惠，请以实际环境为准。 - 三方服务支持申请专属折扣，专属优惠是指伙伴为您申请的优惠折扣。 - 专属折扣申请：单击“优惠折扣”列的“获取专属优惠”链接，跳转至“云商店 -> 商品信息 -> 获取专属优惠”页面，申请专属折扣。 - 专属折扣生效：“推理定价”列展示折扣价格，“优惠折扣”列展示折扣的到期时间。
模型限流	当前账号下，访问同一模型下所有服务的总额度。 - TPM：每分钟处理的Tokens数（输入+输出）。 - RPM：每分钟处理的请求数。
调用统计	单击图标，跳转至“服务调用详情”页面，查看预置服务在指定时间段内的调用数据和监控指标详情。更多信息，请参见在MaaS查看在线推理的调用数据和监控指标。
操作	预置服务支持的相关操作。 - 关闭服务：该按钮置灰，表示暂不支持关闭服务，未使用服务时不会产生费用。 - API调用：参见调用预置服务。 - 在线体验：参见在线体验预置服务。 - 内容安全护栏：仅客户等级为V2及以上，支持开启或关闭内容安全护栏。开通预置服务后，内容安全护栏默认开启。关于客户等级的更多信息，请参见客户等级体系和如何查看客户等级。 在模型推理链路中，开启内容安全护栏，提供额外拦截输入输出有害内容的能力，可能导致时延增加；不开启则使用模型原生安全能力。 - 开启内容安全护栏：单击“更多 > 开启内容安全护栏”，在“开启内容安全护栏”对话框，阅读开启须知，单击“确定”。 - 关闭内容安全护栏：单击“更多 > 关闭内容安全护栏”，在“关闭内容安全护栏”对话框，阅读关闭须知，勾选“我已阅读并同意上述说明，及《内容审核免责声明》”，单击“确定”。 - 开启显式水印：参见开启/关闭显式水印。 - 开启隐式水印：参见开启/关闭隐式水印。

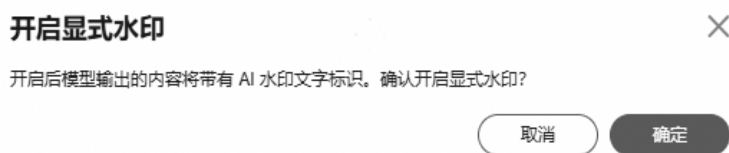
开启/关闭显式水印

对于“视频生成”或“图片生成”类型的预置服务，MaaS支持开启或关闭显式水印。开通预置服务后，显式水印默认关闭。

开启显式水印，模型输出的内容将带有AI水印文字标识。您可以通过以下方式进行开启或关闭。

- 开启显式水印：在已开通的预置服务的“操作”列，单击“更多 > 开启显式水印”，在“开启显式水印”对话框，单击“确定”。

图 3-6 开启显式水印



- 关闭显式水印：在已开通的预置服务的“操作”列，单击“更多 > 关闭显式水印”，在“关闭显式水印”对话框，单击“确定”。

开启/关闭隐式水印

对于“视频生成”或“图片生成”类型的预置服务，MaaS支持开启或关闭隐式水印。开通预置服务后，隐式水印默认关闭。

开启隐式水印，将依据您主动提供的企业主体信息和该主体对应的统一社会信用代码为您生成文件元数据隐式标识中的Label、ContentProducer和ContentPropagator字段信息。您可以通过以下方式进行开启或关闭。

- 仅平台服务支持隐式水印。
- 开启隐式水印：在已开通的预置服务的“操作”列，单击“更多 > 开启隐式水印”，按需勾选“内容传播服务提供者”，输入“社会统一信用代码”、“主体名称”，查看隐式水印使用说明，勾选“我已阅读并同意上述使用说明”，单击“确定”。

图 3-7 开启隐式水印

开启隐式水印

☒ 生成合成服务提供者

☐ 内容传播服务提供者

社会统一信用代码

请输入

18位，支持数字或字母

主体名称

请输入

不超过200个字，仅支持中文

隐式标识 Metadata: ProduceID 与 PropagatorID 默认使用接口返回头中的 x-span-id，无需手动填写。
声明：MaaS 仅提供技术工具，不会主动审核您所标识内容的完整性、准确性及合规性，请确保填写的信息真实有效。

MaaS大模型即服务平台控制台隐式水印使用说明

- 用户您好，根据《生成式人工智能服务管理暂行办法》、《人工智能生成内容标识办法》等相关法律法规的规定，经过您的判断，如您需要对人工智能生成内容添加文件元数据隐式标识，您可开启本功能。
- 功能说明：本功能将依据您主动提供的企业主体信息和该主体对应的统一社会信用代码为您生成文件元数据隐式标识中的Label、ContentProducer和ContentPropagator字段信息。
- 您理解并确认，元数据隐式标识相关技术标准规定的其他字段信息，需要您自行补充。
- 您理解并同意，由于文件元数据隐式标识存在被修改的可能性，您需要对相关标识内容的完整性、准确性和合规性承担全部责任。MaaS仅为您提供技术工具，MaaS不会主动审核您标识内容的完整性和准确性。

☐ 我已阅读并同意上述使用说明

取消

确定

- 关闭隐式水印：在已开通的预置服务的“操作”列，单击“更多 > 关闭隐式水印”，在“关闭隐式水印”对话框，单击“确定”。

调用预置服务

1. 在已开通的预置服务右侧，单击“API调用”，选择服务版本。
2. 在“API调用说明”面板，按照页面提示调用预置服务。
 - 关于如何获取API Key，请参见[管理API Key](#)。
 - 不同模型类型的API参数不同，详情请参见[模型API调用规范](#)。
 - 在调用预置服务时，如果出现报错，请参考[错误码](#)定位问题。

在线体验预置服务

在已开通的预置服务右侧，单击“更多 > 在线体验”，选择服务版本，跳转至模型对应的体验页面，进行在线体验。更多信息，请参见[MaaS在线体验](#)。

如果没有“在线体验”按钮，表示预置服务不支持该操作。您可以使用API调用该预置服务，详情请参见[调用预置服务](#)。

流控规则说明

为了保证用户调用模型的公平性，MaaS设置了基础限流。如果超出限制，API请求将会失败，需等到解除限流条件时再次调用。

- TPM（Tokens Per Minute）：每分钟处理的Tokens数（输入+输出）。
- RPM（Requests Per Minute）：每分钟处理的请求数。

建议您均匀地发送API请求，避免在短时间内发送大量请求。限流策略可能按照秒级QPS（RPM/60）限制，即使调用总量未达到每分钟上限，短时间内发送大量请求也可能触发限流，导致请求失败并返回错误码ModelArts.81101（Too Many Requests）。关于错误码详情，请参见[错误码](#)。

常见问题

- **有计费示例吗？**
计费项和计费示例请参考[在线推理模型服务价格](#)。
- **开通预置服务后，可以关闭吗？**
暂不支持关闭预置服务，未使用服务时不会产生费用。
- **使用预置服务，付费状态显示冻结，如何处理？**
处理“冻结”状态的方法，请参见[使用预置服务，付费状态显示冻结，如何处理？](#)。
- **使用预置服务，付费状态显示失效，如何处理？**
付费状态显示失效，表明资源因账户欠费被冻结，且已超出资源保留期，您可以重新开通预置服务。

4 使用自定义接入点

MaaS支持用户创建自定义接入点，通过自定义接入点名称进行模型调用（model参数设置），实现不同业务场景或模型版本的分流与精细化管理。

在使用大语言模型进行推理服务时，推理接入点（Endpoint）是模型调用的关键入口。开发者只需通过指定的ID即可向大模型发起API推理请求，实现快速接入与灵活调用。

然而在企业 and 开发者的AI应用开发与运营过程中，面临着推理服务调用管理无序、流量控制困难、成本核算模糊等问题。多个业务线共用同一推理服务，导致资源争抢、服务性能不稳定，同时缺乏有效的调用限制手段，难以追溯各业务模块的资源消耗情况。

MaaS支持自定义接入点功能，通过创建独立的调用入口，允许用户设置限流规则，并基于自定义接入点名称实现费用的精准统计，帮助用户高效管理推理服务资源，优化使用成本。

约束限制

- 该功能仅支持“西南-贵阳一”区域。
- 在“自定义接入点”页面的左上方显示可创建自定义接入点的剩余数量。



- 同一账户下不允许存在同名的自定义接入点。已删除的接入点名称不允许新建时使用。
- 自定义接入点创建后，不支持修改模型服务。
- 创建的自定义接入点需遵循平台相关的规则和规范，不得进行违规调用。

计费说明

自定义接入点功能本身不收费。调用模型服务或使用资源可能会产生费用。您可以通过接入点名称在费用中心查询服务使用账单。

- 调用MaaS自定义接入点：由MaaS计费。支持按照Token计费。关于计费详情，请参见[在线推理模型服务价格](#)。

- 调用ModelArts在线服务：由ModelArts计费。在ModelArts进行服务部署时，会产生计算资源和存储资源的累计值计费。关于计费详情，请参见[推理部署（新版）计费项](#)。

前提条件

在ModelArts部署模型为在线服务时，配置要求如下：

- 在“网络配置”中，“认证方式”选择“无认证”，“更多配置”中打开“外网访问”和“内网接入免审批”开关。

创建自定义接入点

- 登录[MaaS控制台](#)，在顶部导航栏中选择目标区域。
- 在左侧导航栏，选择“模型推理 > 在线推理”，单击“自定义接入点”页签，页面左上方显示可创建自定义接入点的剩余数量。



- 单击“自定义接入点”页签，在页面右上角单击“创建自定义接入点”。
- 在“创建自定义接入点”页面，配置相关参数。

表 4-1 创建自定义接入点参数说明

参数		说明
基础信息	名称	自定义接入点的名称。自定义接入点名称具有 唯一性 ，不能重复，不支持特殊字符。支持输入1~100个字符，以数字、字母或中文字符开头，只包含中文、大小写字母、数字、中划线、下划线和小数点的名称。
	描述	自定义接入点的描述，最多支持256字符。

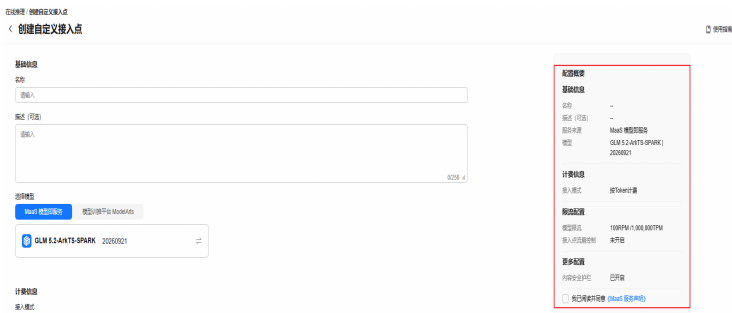
参数		说明
	选择模型	支持选择“MaaS 模型即服务”或“模型训推平台 ModelArts”。 - MaaS模型即服务：MaaS提供的预置模型。 单击“请选择模型”，在“选择模型”对话框，按需选择模型和版本，单击“确定”。预置服务中仅平台服务支持创建自定义接入点。 模型版本右侧会显示支持的接入模式。 图 4-1 模型支持的接入模式 版本 (1) ☐ 20260202 按Token计费 模型单元 版本升级昇腾云6.5.1905版本镜像，基于vllm框架，性能和精度进一步提升。支... - 模型训推平台 ModelArts：新版在线服务，由 ModelArts进行计费。 单击“请选择模型”，在“选择模型服务”对话框，按需选择区域、工作空间和服务，单击“确定”。 说明 - 仅开启MaaS网关的服务可创建接入点，“选择模型服务”对话框仅展示开启MaaS网关的服务。 - 请确保所选服务符合OpenAI标准规范，否则调用可能失败。
计费信息	接入模式	支持“按Token计费”。
限流配置	模型限流	选择预置模型后，会显示当前账号下访问该模型服务的总限流。 - RPM (Requests Per Minute)：每分钟处理的请求数 - TPM (Tokens Per Minute)：每分钟处理的Tokens数（输入+输出）。
	接入点流量控制	手动设置接入点限流额度。如果该账号下访问同一模型的所有接入点限流总和等于该模型的总限流额度，就能有效避免不同接入点之间争夺流量配额。 - 用户可以针对每个接入点设置不同的RPM和TPM流控，但不能超过账号的模型限流值。 - RPM和TPM流控需为正整数。

参数		说明
限流配置	指定流控时段	仅开启“接入点流量控制”后，显示该参数。 开启“指定流控时段”后，可设置接入点每日特定时段的限流额度，其余时段流控保持不变。您可以单击“添加时段”，设置多个时段的限度额度。设置多个时段时，时间范围不能重叠。 设置指定流控时段后，您可以在“自定义接入点”页签的“流量控制”列，单击“查看更多”，查看指定的流控时段信息。 说明 模型限流、接入点流量控制和指定流控时段的生效优先级：指定流控时段>接入点流量控制>模型限流。 - 假设您同时配置了接入点流量控制和指定流控时段，指定流控时段优先生效，不在指定流控时段的流控按照接入点流量控制。 - 假设您只配置了指定流控时段，指定流控时段优先生效，不在指定流控时段的流控按照模型限流。
更多配置	内容安全护栏	仅客户等级为V2及以上，支持开启或关闭内容安全护栏。默认开启内容安全护栏。更多信息，请参见客户等级体系和如何查看客户等级。 在模型推理链路中，开启内容安全护栏，提供额外拦截输入输出有害内容的能力，可能导致时延增加；不开启则使用模型原生安全能力。 关闭内容安全护栏：取消勾选“内容安全护栏”，阅读关闭须知，勾选“我已阅读并同意上述说明，及《内容审核免责声明》”。 自定义接入点创建后，可以在编辑自定义接入点时，开启或关闭内容安全护栏。具体操作，请参见编辑自定义接入点。
	隐式水印	仅“选择模型”选择“MaaS 模型即服务”，且选择“视频生成”或“图片生成”类型的预置服务，显示该参数。 该功能默认关闭。开启后，您需要对人工智能生成合成内容添加文件元数据隐式标识。 - 开启隐式水印：勾选“隐式水印”，按需勾选“内容传播服务提供者”，输入“社会统一信用代码”、“主体名称”，查看隐式水印使用说明，勾选“我已阅读并同意上述使用说明”。 - 关闭隐式水印：取消勾选“隐式水印”。 自定义接入点创建后，可以在编辑自定义接入点时，开启或关闭隐式水印。具体操作，请参见编辑自定义接入点。

参数		说明
	显式水印	仅“选择模型”选择“MaaS 模型即服务”，且选择“视频生成”或“图片生成”类型的预置服务，显示该参数。 该功能默认关闭。开启后，模型输出的内容将带有AI水印文字标识。 - 开启显式水印：勾选“显式水印”。 - 关闭显式水印：取消勾选“显式水印”。 自定义接入点创建后，可以在编辑自定义接入点时，开启或关闭显式水印。具体操作，请参见编辑自定义接入点。

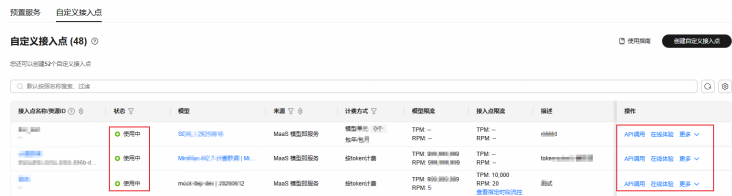
5. 在页面右侧查看配置概要，勾选“我已阅读并同意《MaaS 服务声明》”，在页面下方确认计费无误后，单击“立即创建”。

图 4-2 配置概要



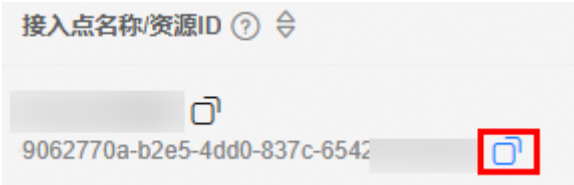
6. 自定义接入点创建成功后，“自定义接入点”页签会显示接入点的相关信息，“状态”会显示为“使用中”，您可以进行调用、在线体验等操作。

图 4-3 自定义接入点创建成功



您还可以在接入点资源ID后，单击图标复制资源ID，在[费用中心](#)的“账单管理 > 流水和明细账单”页面的“明细账单”页签，通过ID查看账单详情。

图 4-4 复制 ID



查看自定义接入点详情

1. 在[MaaS控制台](#)的“模型推理 > 在线推理”页面，单击“自定义接入点”页签，在目标接入点的“接入点名称/资源ID”列，单击目标接入点的接入点名称/资源ID。默认展示“概览”页签，您可以查看到及接入点的详细信息，包括：基础信息、计费信息、限流信息以及更多配置，参数详细信息请参见[表4-2](#)。

表 4-2 自定义接入点参数说明

参数		说明
基础信息	接入点名称	自定义接入点的名称。
	接入点 ID	自定义接入点的ID。
	模型	接入点选择的模型。单击模型名称，您可以查看该模型的信息。
	来源	接入点选择模型的来源。 目前支持“MaaS 模型即服务”或“模型训推平台 ModelArts”。
	model参数	接入点选择模型的model参数。
	创建时间	自定义站点的创建时间。
	描述	自定义接入点的描述。
计费信息	计费方式	自定义接入点的计费方式。目前支持“按Token计费”。详细信息，请查看 表4-1
	推理定价	推理定价单位。单击定价，您可以切换“千tokens”或“百万tokens”。
限流信息	模型限流	仅“选择模型”选择“MaaS 模型即服务”，显示该参数。 选择预置模型后，会显示当前账号下访问该模型服务的总限流。 ● RPM（Requests Per Minute）：每分钟处理的请求数 ● TPM（Tokens Per Minute）：每分钟处理的Tokens数（输入+输出）。
	接入点限流	手动设置接入点限流额度。如果该账号下访问同一模型的所有接入点限流总和等于该模型的总限流额度，就能有效避免不同接入点之间争夺流量配额。 ● 用户可以针对每个接入点设置不同的RPM和TPM流控，但不能超过账号的模型限流值。 ● RPM和TPM流控需为正整数。
更多配置	内容安全护栏	展示内容安全护栏当前状态，包括“已开启”和“未开启”。

参数		说明
	隐式水印	仅“选择模型”选择“MaaS 模型即服务”，且选择“视频生成”或“图片生成”类型的预置服务，显示该参数。 该功能默认关闭。开启后，您需要对人工智能生成合成内容添加文件元数据隐式标识。
	显示水印	仅“选择模型”选择“MaaS 模型即服务”，且选择“视频生成”或“图片生成”类型的预置服务，显示该参数。 该功能默认关闭。开启后，模型输出的内容将带有AI水印文字标识。

- 自定义接入点支持编辑、删除、启用/停用，在线体验以及API调用。
 - 您还可以在接入点资源ID后，单击图标复制资源ID，在[费用中心](#)的“账单管理 > 流水和明细账单”页面的“明细账单”页签，通过ID查看账单详情。
 - 您还可以在model参数后，单击图标复制model参数，在“API调用”页签，用户可以根据不同的model参数进行不同接入点的调用。
2. 单击“API调用说明”页签，查看调用自定义接入点具体调用流程，详情请参见[调用自定义接入点](#)。
3. 单击“监控”页签，查看自定义接入点的查看调用详情。
- “监控”页签按需选择时间范围、时间颗粒维度、调用方式和IP地址。

表 4-3 接入点监控筛选参数说明

参数	说明
时间范围	支持按照今天、昨天、近三天、近7天、近14天、自定义时间段统计接入点的调用数据。 时间范围与时间精度过滤规则： • 时间范围1~2天：支持按分钟、小时进行统计。 • 时间范围为3~7天：支持按小时、天进行统计。 • 时间范围为7天以上：支持按天进行统计。
时间颗粒度	支持按分钟或者按小时统计接入点的调用数据。
调用方式	支持API Key调用和在线体验。 • API Key调用：调用MaaS的预置模型服务时，使用API Key进行鉴权认证，默认为“全部API Key”，您也可以按需勾选API Key。 • 在线体验：在线体验模型服务产生的调用数据。更多信息，请参见MaaS在线体验。

参数	说明
IP地址	已产生调用量的客户端源IP地址（公网IP），来源于APIG日志中的http_x_forwarded_for字段值。当该字段包含多个值时，系统将采用第一个值；当字段值为-时，显示为空字符串。 IP地址默认显示为“全部”，您也可以按需勾选IP地址。

- 用量：跟踪接入点的Tokens使用量，避免超额使用。

表 4-4 用量指标参数说明

参数	说明
已完成任务数量（个）	接入点已完成任务个数。
推理次数(次)	接入点推理次数。
调用次数（次）	接入点调用成功、失败的次数。
调用tokens量（千tokens）	单位时间内接入点的调用总tokens数。
TPM（千tokens/分钟）	TPM（Tokens Per Minute）每分钟处理的Tokens数（输入+输出）。
RPM（次/分钟）	RPM（Requests Per Minute）每分钟处理的请求数。
QPS（次/秒）	QPS（Queries Per Second）表示每秒查询数。当“时间精度”设置为“按分钟”时，实际显示的是该分钟内的QPS峰值。
输入Tokens大小（千tokens）	输入Token长度。 • AVG：输入Token长度的平均值。 • MAX：输入Token长度的最大值。 • P50：50%的输入Token长度低于该值。 • P80：80%的输入Token长度低于该值。 • P90：90%的输入Token长度低于该值。 • P99：99%的输入Token长度低于该值。
输出Tokens大小（千tokens）	输出Token长度。 • AVG：输出Token长度的平均值。 • MAX：输出Token长度的最大值。 • P50：50%的输出Token长度低于该值。 • P80：80%的输出Token长度低于该值。 • P90：90%的输出Token长度低于该值。 • P99：99%的输出Token长度低于该值。

- 性能：支持查看接入点的多种常见性能指标，进行性能优化。

表 4-5 性能指标指标参数说明

参数	说明
平均生成时长（s）	平均生成每张图片或每个视频实际花费的时间。
端到端时延（ms）	单位时间内成功请求的端到端时延。 • AVG：端到端时延的平均值。 • MAX：端到端时延的最大值。 • P50：50%的端到端时延低于该值。 • P80：80%的端到端时延低于该值。 • P90：90%的端到端时延低于该值。 • P99：99%的端到端时延低于该值。
首Token时延（ms）	从接收请求到生成第一个输出Token所需的时间， 仅统计流式响应 。受限于模型版本约束，部分模型版本在非流式场景下不支持该指标展示，请将该服务的模型升级至最新版本后查看。 • AVG：首Token时延的平均值。 • MAX：首Token时延的最大值。 • P50：50%的首Token时延低于该值。 • P80：80%的首Token时延低于该值。 • P90：90%的首Token时延低于该值。 • P99：99%的首Token时延低于该值。
增量Token时延（ms）	生成后续每个输出Token所需的时间间隔， 仅统计流式响应 。受限于模型版本约束，部分模型版本在非流式场景下不支持该指标展示，请将该服务的模型升级至最新版本后查看。 AVG：增量Token时延的平均值。 • MAX：增量Token时延的最大值。 • P50：50%的增量Token时延低于该值。 • P80：80%的增量Token时延低于该值。 • P90：90%的增量Token时延低于该值。 • P99：99%的增量Token时延低于该值。
平均任务处理时长(min)	平均处理任务花费的时间。

- 缓存：支持查看接入点的缓存命中指标，评估系统性能。

表 4-6 缓存指标指标参数说明

参数	说明
缓存命中数 (千tokens)	接入点缓存命中的Tokens数。
缓存命中率 (%)	接入点缓存命中的Tokens数占输入的Tokens数的比例。

- 异常：快速定位特定时间段的调用量激增、异常消耗和调用失败问题。

表 4-7 异常指标指标参数说明

参数	说明
调用失败率 (%)	调用失败次数占调用总次数的比例。
调用失败次数 (次)	各错误码的发生次数。
推理失败率 (%)	推理失败次数占推理总次数的比例。

- 其它相关调用统计的更多信息，请参见[在MaaS查看在线推理的调用数据和监控指标](#)。

在线体验自定义接入点

只有当自定义接入点的“状态”为“使用中”，才能进行在线体验。

1. 在[MaaS控制台](#)的“模型推理 > 在线推理”页面，单击“自定义接入点”页签，在目标接入点的“操作”列，单击“在线体验”。
- 关于在线体验的更多信息，请参见[体验文本对话](#)。

调用自定义接入点

只有当自定义接入点的“状态”为“使用中”，才能被成功调用。服务调用产生的内容由AI生成，不代表MaaS观点，平台不保证其合法性、真实性、准确性，不承担相关法律责任。

1. 在[MaaS控制台](#)的“模型推理 > 在线推理”页面，单击“自定义接入点”页签，在目标接入点的“操作”列，单击“API调用”。
2. 在“API调用说明”页面，按照页面提示获取API Key，复制调用示例并替换接口信息、API Key，进行API调用。
 - 在“自定义接入点”页签的“model参数”列显示的名称，为调用服务时代码的model参数值。用户可以根据不同的model参数进行不同接入点的调用。
 - 关于如何创建API Key，请参见[创建API Key](#)。
 - 关于调用示例的参数说明，请参见[MaaS标准API V2](#)。

管理推理接入点

自定义接入点支持查看调用数据、指标监控，支持编辑、启用/停用等操作。

查看自定义接入点的调用统计

您可以查看自定义接入点在指定时间段内的调用数据和监控指标详情，包括调用次数、调用失败次数、调用总Tokens数等信息，了解服务的使用情况和性能变化，从而更有效地进行模型评估、问题定位、故障排除和性能优化。

1. 在[MaaS控制台](#)的“模型推理 > 在线推理”页面，单击“自定义接入点”页签。
2. 在目标接入点的“调用统计”列，单击图标，跳转至“服务调用详情”页面，查看调用详情。

关于调用统计的更多信息，请参见[在MaaS查看在线推理的调用数据和监控指标](#)。

编辑自定义接入点

您可以按需修改自定义接入点信息，例如描述、限流等。**自定义接入点的模型服务不支持修改。**

1. 在[MaaS控制台](#)的“模型推理 > 在线推理”页面，单击“自定义接入点”页签，在目标接入点的“操作”列，单击“更多 > 编辑”。
2. 在“编辑自定义接入点”面板，按需修改相关参数，单击“更新”。

关于参数说明，请参见[创建自定义接入点参数说明](#)。

停用/启用自定义接入点

当自定义接入点“状态”为“使用中”，可以停用自定义接入点。停用接入点后，该接入点的推理能力将停用，支持重新启用。由于出账存在时延，可能在您停用后仍会收到由该服务产生的账单。

当自定义接入点“状态”为“停用”，可以启用自定义接入点。启用后，调用或体验该接入点时将按照原计费方式产生费用。

1. 在[MaaS控制台](#)的“模型推理 > 在线推理”页面，单击“自定义接入点”页签，按需选择以下操作。
 - 停用自定义接入点
 - i. 在目标接入点的“操作”列，单击“更多 > 停用”。
 - ii. 在“停用自定义接入点”对话框，单击“确定”。停用后，该接入点的状态会显示为“停用”。
 - 启用自定义接入点
 - i. 在目标接入点的“操作”列，单击“更多 > 启用”。
 - ii. 在“启用自定义接入点”对话框，单击“确定”。启用后，该接入点的状态会显示为“使用中”。

删除自定义接入点

当自定义接入点不再需要时，您可以进行删除操作。**删除后，该接入点的推理能力将停用，全部信息将被删除且无法恢复，请谨慎操作。**

由于出账存在时延，可能在您删除后仍会收到由该服务产生的账单。

1. 在[MaaS控制台](#)的“模型推理 > 在线推理”页面，单击“自定义接入点”页签，在目标接入点的“操作”列，单击“更多 > 删除”。

2. 在“删除自定义接入点”对话框，查看删除提示信息，确认无误后输入**DELETE**，单击“确定”。
删除后，“自定义接入点”页签将不再显示该接入点。

常见问题

- **创建的自定义接入点数量达到上限怎么办？**
您可以删除不再使用的接入点，然后新建接入点。
- **如何确定Tokens的消耗数量？**
您可以通过以下两种方式查看Tokens的消耗数量。
 - 通过“调用统计”页面查看模型服务调用的总Tokens数、输入Tokens数、输出Tokens数等信息，详情请参见[查看自定义接入点的调用统计](#)。
 - 在费用中心通过自定义接入点名称查询账单详情（该方式仅支持预置模型接入点）。账单中会显示接入点的输入Tokens数、输出Tokens数等信息。
- **修改自定义接入点的限流设置后，多久会生效？**
修改保存后，限流设置会立即生效，后续调用将按照新规则执行。
- **自定义接入点的付费状态变为失效怎么办？**
自定义接入点的付费状态变为失效，说明资源因账户欠费被冻结，且已超出资源保留期，导致资源被删除，您可以删除此接入点并重新创建。

5 MaaS 在线体验

5.1 在 MaaS 体验文本对话

在使用MaaS进行模型测试时，用户常需快速验证模型的文本对话效果，但传统方法往往因配置复杂和准备时间长而难以满足这一需求。如何在保证准确性的同时，快速体验模型的文本对话效果？MaaS支持使用“预置服务”、“自定义接入点”进行在线文本对话，还支持对比同一文本对话模型的不同版本或不同模型，帮助您快速体验模型的文本对话效果，提高测试效率。

约束限制

该功能仅支持“西南-贵阳一”区域。

计费说明

使用预置服务、使用预置服务创建的自定义接入点进行文本对话，输入和输出会转化为Token计费。关于计费详情，请参见[MaaS文本模型计费项](#)。

前提条件

- 使用预置服务：在“模型推理 > 在线推理 > 预置服务”页签，已开通预置服务。具体操作，请参见[使用预置服务](#)。
- 使用自定义接入点：在“模型推理 > 在线推理 > 自定义接入点”页签，已创建自定义接入点。具体操作，请参见[创建自定义接入点](#)。

在线体验文本对话

1. 登录[MaaS控制台](#)，在顶部导航栏中选择目标区域。
2. 任选以下方式进行模型体验。
 - 方式一
 - i. 在左侧导航栏中，选择“模型体验 > 文本对话”。
 - ii. 在“文本对话”页面左上角单击图标切换模型，在“预置服务”、“自定义接入点”页签，选择要体验的模型服务，单击“确定”。如下图所示：



- “预置服务”页签：选择目标服务及版本进行体验。模型列表中的模型包括：已开通模型和未开通模型。
如果选择已开通模型或者切换版本，单击“确认”，完成选择模型或者切换模型。
如果选择未开通模型，则弹出“开通模型服务”弹框。在弹框中确认“服务名称”、“版本”、“供给方式”、“计费方式”以及“费用评估”等信息后，阅读《MaaS 模型即服务声明》后，勾选“我已阅读并同意上述说明，及《MaaS 模型即服务声明》”后，单击“确认”，完成选择模型或者切换模型。

图 5-1 开通模型服务



- “自定义接入点”页签：选择状态为“使用中”的自定义接入点，单击“确认”，选择自定义接入点完成。
- 方式二
 - i. 在左侧导航栏中，选择“模型推理 > 在线推理”。
 - ii. 任选以下方式进入“文本对话”页面。
 - 在“预置服务”页签，单击操作列的“更多 > 在线体验”，选择版本，进入“文本对话”页面。
 - 在“自定义接入点”页签：单击操作列的“在线体验”，进入“文本对话”页面。
 - 方式三
 - 选择左侧导航栏“模型广场”，进入“模型广场”页面，单击模型卡片上的“在线体验”按钮，直接进入“文本对话”页面。

3. （可选）在“文本对话”页面左上方，单击“模型参数”，弹出“参数设置”页面，如下图所示：

图 5-2 设置推理参数

参数设置

温度/Temperature 0.7

012

核采样/top_p 1

0.10.551

top_k 20

15001000

恢复默认

根据表5-1中的参数说明，您可按需设置推理参数，体验不同的推理效果。如需恢复默认参数，单击“恢复默认”即可。

表 5-1 模型参数设置

参数	说明
温度/Temperature	设置推理温度，用于控制生成文本的随机性和创造性，Temperature数值越大随机性越大。 - 数值较低，输出结果更加集中和确定。 - 数值较高，输出结果更加随机，更有创意性。 取值范围：0~2 默认值：不同模型的默认值不同，请以实际环境为准。
核采样/top_p	设置推理核采样，用于调整输出文本的多样性。top_p数值越大，生成文本的多样性就越高。 - 数值较低，输出可选的tokens类型越少，更有确定性。 - 数值较高，输出可选的tokens类型越多，更有多样性。 取值范围：0.1~1 默认值：不同模型的默认值不同，请以实际环境为准。

参数	说明
top_k	用于控制输出tokens的多样性。top_k值越大输出的tokens类型越丰富。选择在模型的输出结果中选择概率最高的前K个结果。 • 数值较低，输出可选的tokens类型越少，更有确定性。 • 数值较高，输出可选的tokens类型越多，更有多样性。 取值范围：1~1000 默认值：20 详细解释：top_k可以设置保留概率最高的前K个tokens，从中随机抽取一个token作为最终输出。这种方法可以限制输出序列的长度，并仍然保持样本的一定多样性。

4. 在对话框中输入问题或者使用控制台提供的推荐词，查看返回结果，在线体验模型服务。
- 模型输出内容不代表平台观点，平台不保证其合法性、真实性、准确性，不承担相关法律责任。**输入和输出内容已默认开启内容审核。**

图 5-3 体验模型服务



在线体验功能相关操作说明如下：

- 切换模型/版本：在模型名称右侧单击  图标，按需选择模型和版本，单击“确定”。
- 停止模型回答：在模型回答过程中，您可以在对话框单击  图标，停止模型回答，停止后该对话不产生计费。
- 复制回答：对于生成的内容，您可以在模型对话框左下方，单击  图标复制回答。
- 重新生成回答：在模型对话框左下方，单击  图标重新生成回答。
- 开启新对话：在页面右上角，单击“开启新对话”，在对话框中单击“确定”。开启新对话将清除全部对话，无法保留对话记录。
- API调用：在页面右上角，单击“API调用”，可以查看模型的调用信息，请按照页面提示进行调用。详情请参考[文本生成](#)。

模型对比

MaaS支持对比同一文本对话模型的不同版本或不同模型，最多可同时对比三个模型/版本。

1. 在“文本对话”页面右上角，单击“模型对比”。
- “模型对比”按钮中会显示当前的模型对比个数，2/3表示当前已对比两个模型。再次单击“模型对比”，可增加模型对比个数。当模型对比数为3/3时，达到模型对比上限，“模型对比”按钮会置灰。

图 5-4 对比两个模型

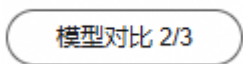


图 5-5 对比三个模型



- 切换模型/版本：MaaS会自动选择模型进行对比，您可以在模型名称右侧单击  图标，按需选择模型和版本，单击“确定”。
- API调用：在模型名称右侧，单击  图标，可以查看模型的调用信息，请按照页面提示进行调用。关于API Key参数说明，请参见[创建API Key](#)；关于API调用的参数说明，请参见[对话Chat/Post](#)。
- 修改推理参数：在模型名称右侧，单击  图标，可以对温度/Temperature、核采样/top_p、top_k参数进行修改，获得更好的模型推理效果。关于参数说明，请参见[表5-1](#)。
- 关闭/新增对比模型：在模型名称右侧，单击  图标，可以关闭该对比模型；单击“模型对比”，可以新增对比模型。

图 5-6 模型对比相关功能



2. 在对话框中输入问题或者使用控制台提供的推荐词，查看模型对比的效果。

图 5-7 查看模型对比效果



如果您不想使用模型对比功能，可以在页面右上角，单击“退出对比”。

5.2 在 MaaS 体验图像理解

在日常工作中，当您需要快速理解图像内容时（如分析产品设计图、解读活动现场照片等），通常会遇到人工处理耗时耗力、细节遗漏和专业门槛等问题。如何高效解决这些问题？MaaS提供图像理解功能，帮助您快速准确地分析图像内容，提高工作效率。

约束限制

- 该功能仅支持“西南-贵阳一”区域。
- 单次对话一次只能发送一张图片，且必须携带文字。图片大小不能超过10MB，图片格式为jpg、jpeg、png或webp。

计费说明

使用预置服务或者使用预置服务创建的自定义接入点：统计模型进行实时推理服务时消耗的Token数量，按Token数量付费，详情请参见[图像理解模型计费](#)。

前提条件

- 使用预置服务：在“模型推理 > 在线推理 > 预置服务”页签，已开通图像理解模型的预置服务。具体操作，请参见[使用预置服务](#)。
- 使用自定义接入点：在“模型推理 > 在线推理 > 自定义接入点”页签，已创建图像理解模型的自定义接入点。具体操作，请参见[创建自定义接入点](#)。

体验图像理解

- 登录[MaaS控制台](#)，在顶部导航栏中选择目标区域。
- 任选以下方式体验图像理解。
 - 方式一
 - 在左侧导航栏中，选择“模型体验 > 图像理解”。
 - 在“图像理解”页面左上角单击图标切换模型，在“预置服务”、“自定义接入点”页签，选择要体验的模型服务，单击“确定”。如下图所示：



- “预置服务”页签：选择目标服务及版本进行体验。模型列表中的模型包括：已开通模型和未开通模型。
如果选择已开通模型或者切换版本，单击“确认”，完成选择模型或者切换模型。
如果选择未开通模型，则弹出“开通模型服务”弹框。在弹框中确认“服务名称”、“版本”、“供给方式”、“计费方式”以及“费用评估”等信息后，阅读[《MaaS 模型即服务声明》](#)后，勾选

“我已阅读并同意上述说明，及《MaaS 模型即服务声明》”后，单击“确认”，完成选择模型或者切换模型。

图 5-8 开通模型服务



- “自定义接入点”页签：选择状态为“使用中”的自定义接入点，单击“确认”，选择自定义接入点完成。
- 方式二
 - i. 在左侧导航栏中，选择“模型推理 > 在线推理”。
 - ii. 任选以下方式进入“图像理解”页面。
 - 在“预置服务”页签，单击图像理解模型服务操作列的“更多 > 在线体验”，按需选择版本，进入“图像理解”页面。
 - 在“自定义接入点”页签：单击图像理解模型服务操作列的“在线体验”，进入“图像理解”页面。
- 方式三
 - 选择左侧导航栏“模型广场”，进入“模型广场”页面，单击模型卡片上的“在线体验”按钮，直接进入“图像理解”页面。
- 3. （可选）在“图像理解”页面左上方，单击“模型参数”，弹出“参数设置”页面，如下图所示：

图 5-9 设置推理参数

参数设置

温度/Temperature ⓘ

0.7

012

核采样/top_p ⓘ

1

0.10.551

top_k ⓘ

20

15001000

恢复默认

根据表5-2中的参数说明，您可按需设置推理参数，体验不同的推理效果。如需恢复默认参数，单击“恢复默认”即可。

表 5-2 模型参数设置

参数	说明
温度/Temperature	设置推理温度，用于控制生成文本的随机性和创造性，Temperature数值越大随机性越大。 - 数值较低，输出结果更加集中和确定。 - 数值较高，输出结果更加随机，更有创意性。 取值范围：0~2 默认值：不同模型的默认值不同，请以实际环境为准。
核采样/top_p	设置推理核采样，用于调整输出文本的多样性。top_p数值越大，生成文本的多样性就越高。 - 数值较低，输出可选的tokens类型越少，更有确定性。 - 数值较高，输出可选的tokens类型越多，更有多样性。 取值范围：0.1~1 默认值：不同模型的默认值不同，请以实际环境为准。

参数	说明
top_k	用于控制输出tokens的多样性。top_k值越大输出的tokens类型越丰富。选择在模型的输出结果中选择概率最高的前K个结果。 • 数值较低，输出可选的tokens类型越少，更有确定性。 • 数值较高，输出可选的tokens类型越多，更有多样性。 取值范围：1~1000 默认值：20 详细解释：top_k可以设置保留概率最高的前K个tokens，从中随机抽取一个token作为最终输出。这种方法可以限制输出序列的长度，并仍然保持样本的一定多样性。

4. 在“图像理解”页面下方，单击图标，上传图片（新上传图片将替换原有图片）并输入问题，或者直接使用控制台提供的示例图片，然后单击图标发送。

图 5-10 图像理解示例



在模型回答过程中，您可以在对话框单击图标，停止模型回答，停止后该对话不产生计费。当模型回答后，您可以单击模型对话框中的图标复制回答，也可以单击图标重新生成回答。

图像理解支持多轮对话。如果您需要开启新的对话，可以在页面右上角单击“开启新对话”，在“开启新对话”对话框中单击“确定”。开启新对话将清除全部对话，无法保留对话记录。

5.3 在 MaaS 体验图片生成

在日常的创意设计和内容创作过程中，用户经常需要根据特定的文字描述快速生成对应的图片，以辅助创意构思或直接用于项目中。然而，传统的图片生成方式往往需要依赖专业的图片处理软件，不仅操作复杂，而且成本较高。

MaaS提供图片生成在线体验功能，用户只需选择合适的模型，按需配置提示词、图片尺寸等信息，即可快速生成所需图片。图片生成后还可以直接下载使用，极大地简化了图片生成流程，提升了创作效率。

约束限制

- 该功能仅支持“西南-贵阳一”区域。
- 一次只能生成一张图片。

计费说明

不同模型的计费方式不同，计费详情请参见[图片生成模型计费](#)。

前提条件

- 使用预置服务：已在“模型推理 > 在线推理 > 预置服务”页签开通Qwen_image模型的预置服务。具体操作，请参见[使用预置服务](#)。
- 使用自定义接入点：已在“模型推理 > 在线推理 > 自定义接入点”页签为Qwen_image模型创建了自定义接入点。具体操作，请参见[使用自定义接入点](#)。

体验图片生成

1. 登录[MaaS控制台](#)，在顶部导航栏中选择目标区域。
2. 任选以下方式体验图片生成。
 - 方式一
 - i. 在左侧导航栏中，选择“模型体验 > 图片生成”。
 - ii. 在“图像生成”页面，单击“模型服务”右侧的图标切换模型，在“预置服务”或“自定义接入点”页签，选择要体验的模型服务，单击“确定”。如下图所示：



- “预置服务”页签：选择目标服务及版本进行体验。模型列表中的模型包括：已开通模型和未开通模型。
如果选择已开通模型或者切换版本，单击“确认”，完成选择模型或者切换模型。
如果选择未开通模型，则弹出“开通模型服务”弹框。在弹框中确认“服务名称”、“版本”、“供给方式”、“计费方式”以及“费用评估”等信息后，阅读[《MaaS 模型即服务声明》](#)后，勾选“我已阅读并同意上述说明，及《MaaS 模型即服务声明》”后，单击“确认”，完成选择模型或者切换模型。

图 5-11 开通模型服务

开通模型服务

- 预置服务为您提供付费API服务，开通后您可使用服务的所有版本。
- 未使用时不会产生费用，暂不支持关闭服务。
- 后续您还可以在[在线推理-预置服务](#)页开通和管理全部模型服务。

服务名称

deepseek-v4.1-

版本

DeepSeek-V4.1

供给方式

平台服务

计费方式

按token计费

费用预估

输入

¥(■■■■ - ■■■■) /百万tokens

输出

¥■■■■ /百万tokens

☒ 我已阅读并同意上述说明，及《[MaaS 模型即服务声明](#)》

取消

确认开通

- “自定义接入点”页签：选择状态为“使用中”的自定义接入点，单击“确认”，选择自定义接入点完成。
 - 方式二
 - i. 在左侧导航栏中，选择“模型推理 > 在线推理”。
 - ii. 在“在线推理”页面，任选以下方式进入“图片生成”页面。
 - 在“预置服务”页签，单击图片生成模型服务操作列的“更多 > 在线体验”，选择目标版本，进入“图片生成”页面。
 - 在“自定义接入点”页签，单击图片生成模型服务操作列的“在线体验”，进入“图片生成”页面。
 - 方式三
 - 选择左侧导航栏“模型广场”，进入“模型广场”页面，单击模型卡片上的“在线体验”按钮，直接进入“图片生成”页面。
3. 在“图片生成”页面的“参数设置”区域，设置相关参数，然后单击“立即生成”。

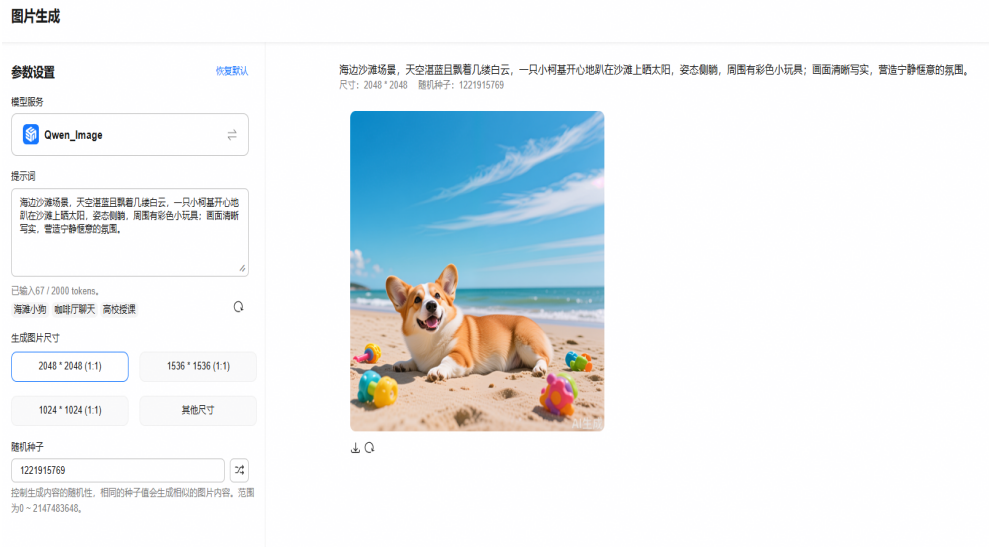
表 5-3 图片生成参数说明

参数	说明
模型服务	单击“选择模型”，在“选择模型服务”对话框，按需选择预置服务或自定义接入点，单击“确定”。
提示词	请描述所需生成图片的场景，或使用控制台提供的推荐内容。单击  图标，可以更换控制台提供的推荐内容。 英文单词、中文字符和标点符号各计为一个Token。最多支持2000个Tokens。

参数	说明
生成图片尺寸	按需选择生成图片的尺寸或自定义生成图片的尺寸。 自定义生成图片的尺寸：单击“其他尺寸”，输入生成图片的宽和高，输入范围为512~3072，建议输入8的倍数。如果输入非8的倍数，将会自动进行匹配。
随机种子	控制生成内容的随机性，相同的种子值会生成类似的图片内容，取值范围为[0,2147483648]。 您可以按需输入，或者单击  图标随机填入数值。

图片生成后，您可以单击图片，查看图片的放大效果，也可以在图片下方单击图标下载图片（格式为png），或者单击图标重新生成图片。

图 5-12 生成图片示例



相关文档

您可以调用API生成图像，详情请参见[图片生成](#)。

5.4 在移动端体验 MaaS 文本对话

MaaS提供了移动端的文本对话功能，帮助您快速在移动端华为云App体验GLM-5.1模型的文本对话效果。

计费说明

使用已开通的预置服务将消耗Token，费用请以实际发生为准。关于如何计费，请参见[计费说明](#)。

前提条件

已注册华为云账号，并进行实名认证。具体操作，请参见[注册华为账号并开通华为云](#)和[实名认证](#)。

场景一：新用户使用网址体验 MaaS

1. 打开[MaaS](#)，跳转至模型服务体验页面。
2. 在模型服务体验页面，将出现“MaaS 模型即服务声明”对话框，请查看服务声明内容并单击“同意”。
3. 在模型服务体验页面顶部，单击“选择模型服务”，在模型服务右侧单击“开通”。
4. 在“开通预置服务”页面，选择需要开通的预置服务，单击“立即开通”。
开通预置服务后，该服务会显示“已开通”。预置服务将为您提供商用级别的API推理服务。暂不支持关闭预置服务，未使用预置服务时不会产生费用。关于如何计费，请参见[计费说明](#)。
5. 在“预置服务”页签，单击已开通的预置服务，进行文本对话。

您可以在右上角单击  图标，按需拖动调整推理参数以调整输出效果。单击“恢复默认”可以将参数值调回默认值。

表 5-4 参数设置

参数	说明
温度/Temperature	设置推理温度。 • 数值较高，输出结果更加随机。 • 数值较低，输出结果更加集中和确定。 取值范围：0~2 默认值：0.7
核采样/top_p	设置推理核采样。调整输出文本的多样性，数值越大，生成文本的多样性就越高。 取值范围：0.1~1 默认值：1
top_k	选择在模型的输出结果中选择概率最高的前K个结果。 取值范围：1~1000 默认值：20

场景二：新用户使用华为云 App 体验 MaaS

1. 下载华为云App。具体信息，请参见[华为云App介绍](#)。
2. 打开华为云App，在底部导航栏单击“控制台”，在搜索框输入“maas”并单击“搜索”，单击“MaaS模型即服务”卡片。

图 5-13 打开 MaaS 产品



3. 在MaaS平台，单击“控制台”，跳转至模型服务体验页面。

图 5-14 MaaS 控制台



4. 在模型服务体验页面，将出现“MaaS 模型即服务声明”对话框，请查看服务声明内容并单击“同意”。
5. 在模型服务体验页面顶部，单击“选择模型服务”，单击“预置服务”页签，在模型服务右侧单击“开通”。
6. 在“开通预置服务”页面，选择需要开通的预置服务，单击“立即开通”。

开通预置服务后，将为您提供商用级别的API推理服务。暂不支持关闭预置服务，未使用预置服务时不会产生费用。关于如何计费，请参见[计费说明](#)。

7. 在“预置服务”页签，单击已开通的预置服务，进行文本对话。

您可以在右上角单击参数设置图标，按需拖动调整推理参数。单击“恢复默认”可以将参数值调回默认值。关于参数说明，请参见[表5-4](#)。

场景三：老用户使用网址体验 MaaS

1. 打开[MaaS](#)，跳转至模型服务体验页面。

模型服务体验页面会默认显示您已开通的预置服务。

2. 在模型服务体验页面，直接进行文本对话，或者在页面上方单击模型服务名称，按需切换预置模型服务进行文本对话。

关于如何开通预置服务，请参见[开通预置服务](#)。

您可以在右上角单击  图标，按需拖动调整推理参数以调整输出效果。单击“恢复默认”可以将参数值调回默认值。关于参数说明，请参见[表5-4](#)。

场景四：老用户使用华为云 App 体验 MaaS

1. 打开华为云App，在底部导航栏单击“控制台”，在搜索框输入“maas”并单击“搜索”，单击“MaaS模型即服务”卡片。

图 5-15 打开 MaaS 产品



2. 在MaaS平台，单击“控制台”，跳转至模型服务体验页面。

模型服务体验页面会默认显示您已开通的预置服务。

3. 在模型服务体验页面，直接进行文本对话，或者在页面上方单击模型服务名称，按需切换预置模型服务进行文本对话。

关于如何开通预置服务，请参见[开通预置服务](#)。

您可以在右上角单击  图标，按需拖动调整推理参数以调整输出效果。单击“恢复默认”可以将参数值调回默认值。关于参数说明，请参见[表5-4](#)。