

# 高性能应用服务 HAI

## 快速入门



腾讯云

## 【 版权声明 】

©2013–2026 腾讯云版权所有

本文档（含所有文字、数据、图片等内容）完整的著作权归腾讯云计算（北京）有限责任公司单独所有，未经腾讯云事先明确书面许可，任何主体不得以任何形式复制、修改、使用、抄袭、传播本文档全部或部分内容。前述行为构成对腾讯云著作权的侵犯，腾讯云将依法采取措施追究法律责任。

## 【 商标声明 】



及其它腾讯云服务相关的商标均为腾讯云计算（北京）有限责任公司及其关联公司所有。本文档涉及的第三方主体的商标，依法由权利人所有。未经腾讯云及有关权利人书面许可，任何主体不得以任何方式对前述商标进行使用、复制、修改、传播、抄录等行为，否则将构成对腾讯云及有关权利人商标权的侵犯，腾讯云将依法采取措施追究法律责任。

## 【 服务声明 】

本文档意在向您介绍腾讯云全部或部分产品、服务的当时的相关概况，部分产品、服务的内容可能不时有所调整。您所购买的腾讯云产品、服务的种类、服务标准等应由您与腾讯云之间的商业合同约定，除非双方另有约定，否则，腾讯云对本文档内容不做任何明示或默示的承诺或保证。

## 【 联系我们 】

我们致力于为您提供个性化的售前购买咨询服务，及相应的技术售后服务，任何问题请联系 4009100100或 95716。

# 文档目录

## 快速入门

通过 HAI 可信集群快速部署模型服务

通过高性能应用服务 HAI 一键创建应用

# 快速入门

## 通过 HAI 可信集群快速部署模型服务

最近更新时间：2026-03-17 16:44:32

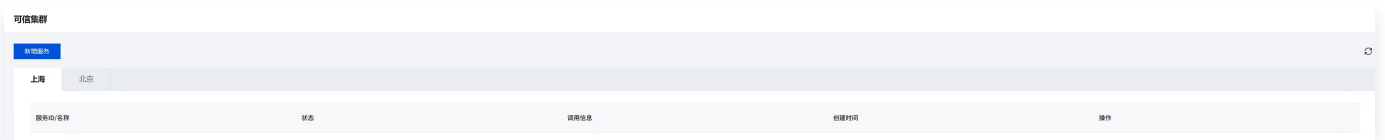
### 步骤1：购买前条件

在购买HAI可信集群资源前，请确保已满足以下条件

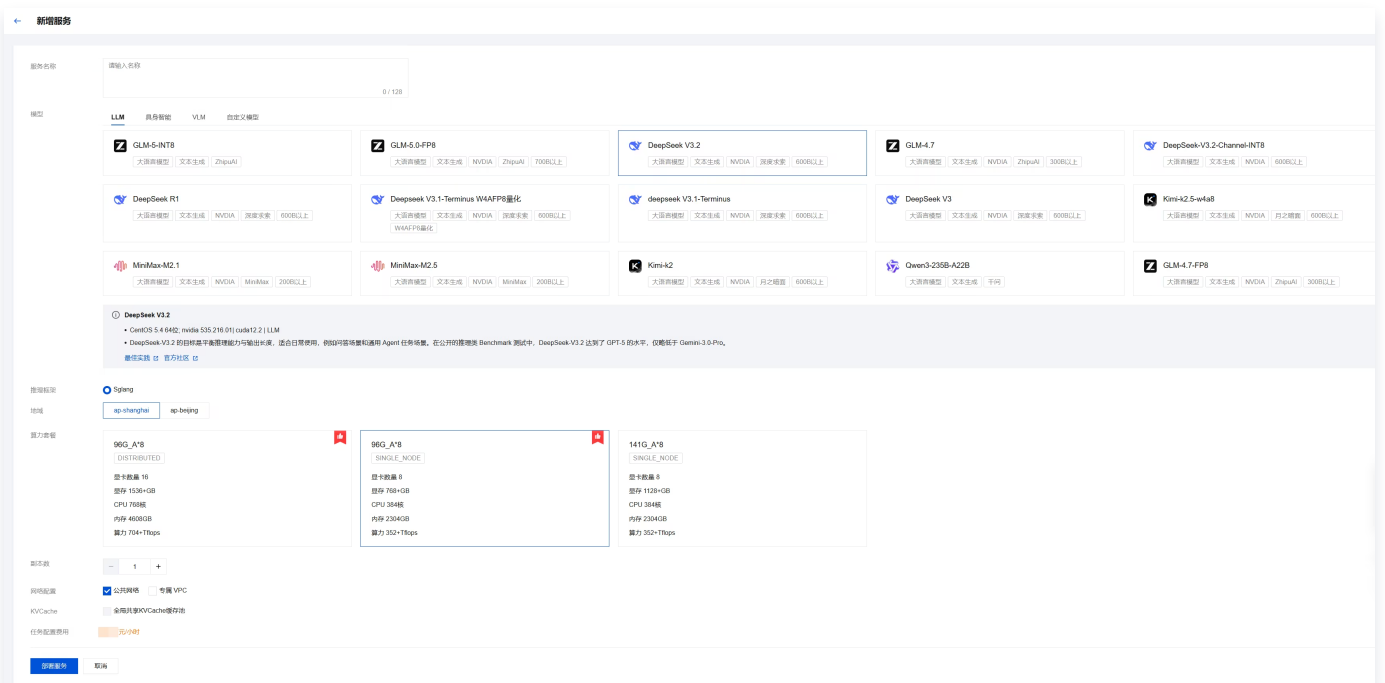
- **已注册腾讯云账号，并完成实名认证。**  
如果您已在腾讯云注册和实名认证，可跳过此步骤。
- **确认账户余额充足。**  
HAI 可信集群资源按量计费，购买前请确保账户余额充足。具体操作请参见 [在线充值](#) 文档。
- **如您是传统账户，需先进行升级。**  
如您当前使用的是传统账户，需先完成账户升级后方可购买。  
账户类型判断及升级方式请参考 [账户类型说明](#)。

### 步骤2：创建 HAI 可信集群

1. 登录 [HAI 可信集群控制台](#)。



2. 单击**新增服务**，进入 [HAI 可信集群购买页面](#)，按照页面指引，完成集群配置。



- **服务名称**: 自定义服务名称, 若不填则默认使用服务的实例 ID 替代。
- **模型**: 选择您想要部署的模型。单击应用后可预览应用环境配置详情及应用介绍信息。
- **地域**: 建议选择靠近目标客户的地域, 降低网络延迟、提高您的客户的访问速度。
- **算力套餐**: 您可根据自己所需的配置进行选择。算力套餐对应的显存、算力、CPU、内存信息在算力方案卡片进行展示。
- **副本数**: 默认单副本。
- **网络配置**: 设置服务访问网络方式。
  - **公共网络**: 适用于快速对外提供服务。
  - **专属 VPC**: 适用于对网络隔离、安全性和内网互通有要求的业务场景。
- **KVCache**: 用于配置是否启用全局共享 KVCache 缓存池。开启后可提升上下文复用效率, 优化推理时延。

**说明:**

该能力正在逐步开放中, 具体支持范围请以控制台展示为准。

- **任务配置费用**: 展示当前配置下的预估费用, 按小时计费, 并随配置项调整实时变化。

3. 单击**部署服务**, 并根据页面提示完成支付。

当您付费完成后, 即完成了该模型的部署服务。接下来, 您可以通过创建资源后生成的调用地址以及您的 token, 调用您的模型服务。详情请参见 [获取可信集群服务调用信息](#)。

# 通过高性能应用服务 HAI 一键创建应用

最近更新时间：2024-08-09 17:35:11

## 步骤1：注册和充值

1. [注册腾讯云账号](#)，并进行实名认证。

如果您已在腾讯云注册，可忽略此步骤。

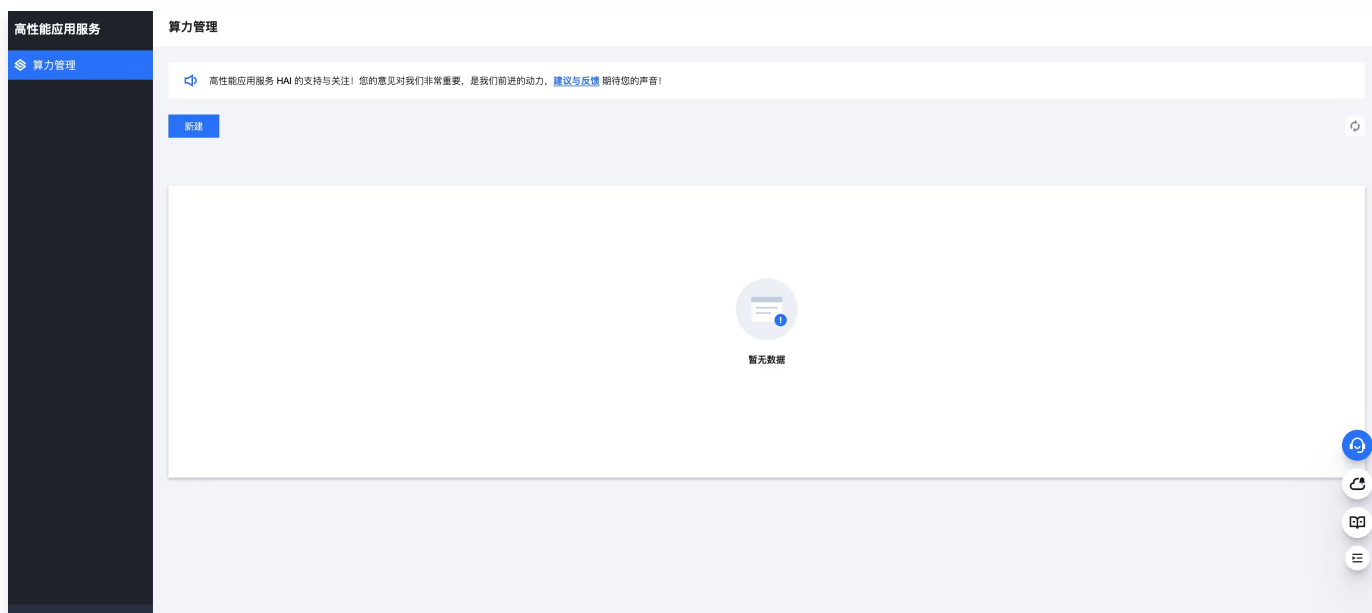
2. [在线充值](#)。

高性能应用服务 HAI 器以按量计费模式售卖，购买前，需要在账号中进行充值。具体操作请参见 [在线充值](#) 文档。

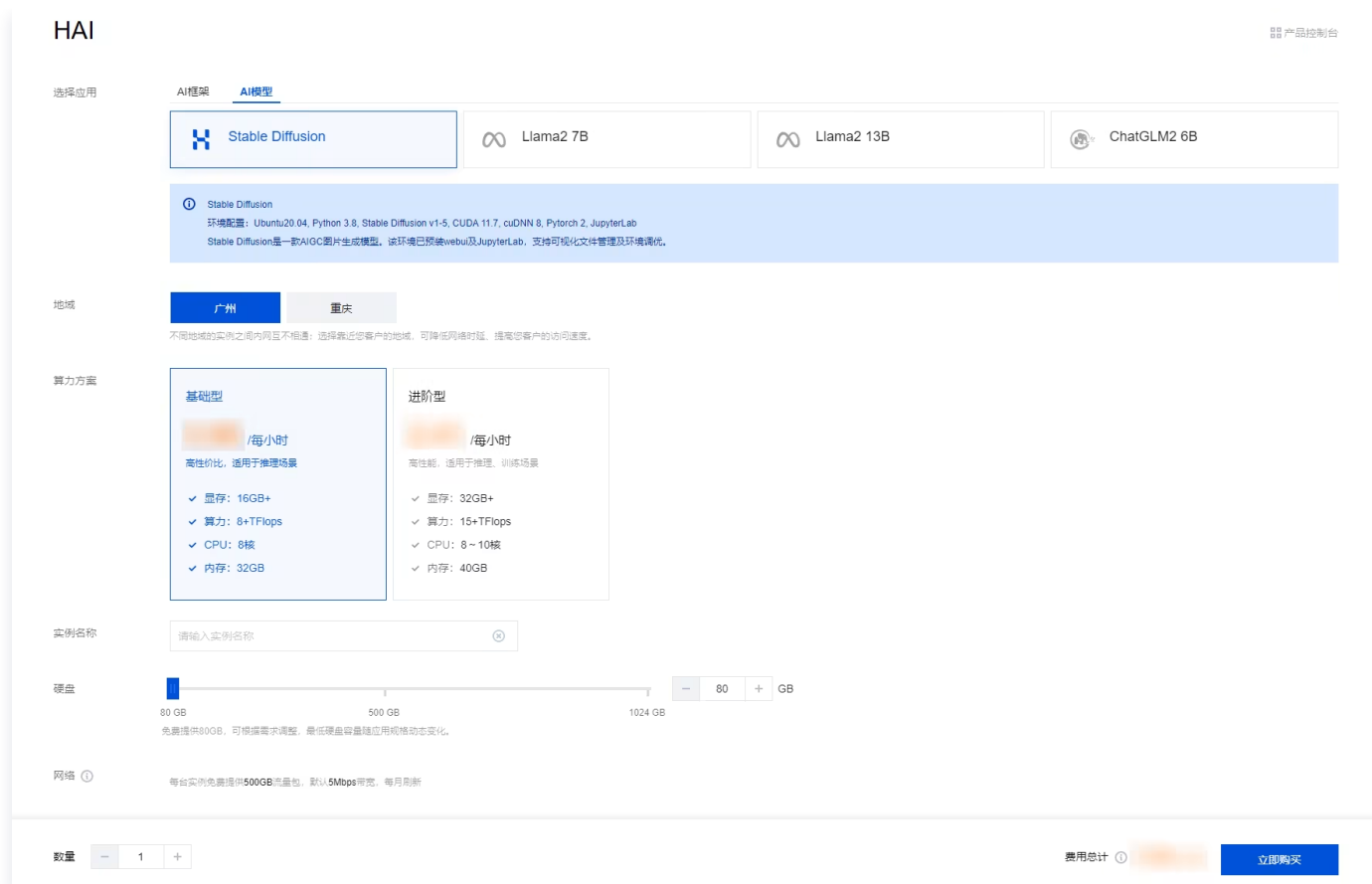
3. 如您是传统账户，需先进行升级。判断账户类型或了解升级方法，可参见 [账户类型说明](#)。

## 步骤2：创建高性能应用服务 HAI

1. 登录 [高性能应用服务 HAI 控制台](#)。



2. 单击新建，进入 [高性能应用服务 HAI 购买页面](#)。



- **选择应用：**目前提供 AI 框架、AI 模型两类应用。单击应用后可预览应用环境配置详情及应用介绍信息。
- **地域：**建议选择靠近目标客户的地域，降低网络延迟、提高您的客户的访问速度。
- **算力方案：**支持基础型及进阶型两类算力方案。用户可根据自己所需的配置进行选择。算力套餐对应的显存、算力、CPU、内存信息在算力方案卡片进行展示。
- **实例名称：**自定义实例名称，若不填则默认使用实例 ID 替代。
- **硬盘：**默认提供 80GB 免费空间，可根据实际使用需求进行调整。
- **网络：**每台实例每月免费提供 500GB 流量包，默认 10Mbps 带宽，每月刷新。
- **购买数量：**默认1台。

### 3. 单击立即购买。

4. 核对配置信息后，单击**提交订单**，并根据页面提示完成支付。

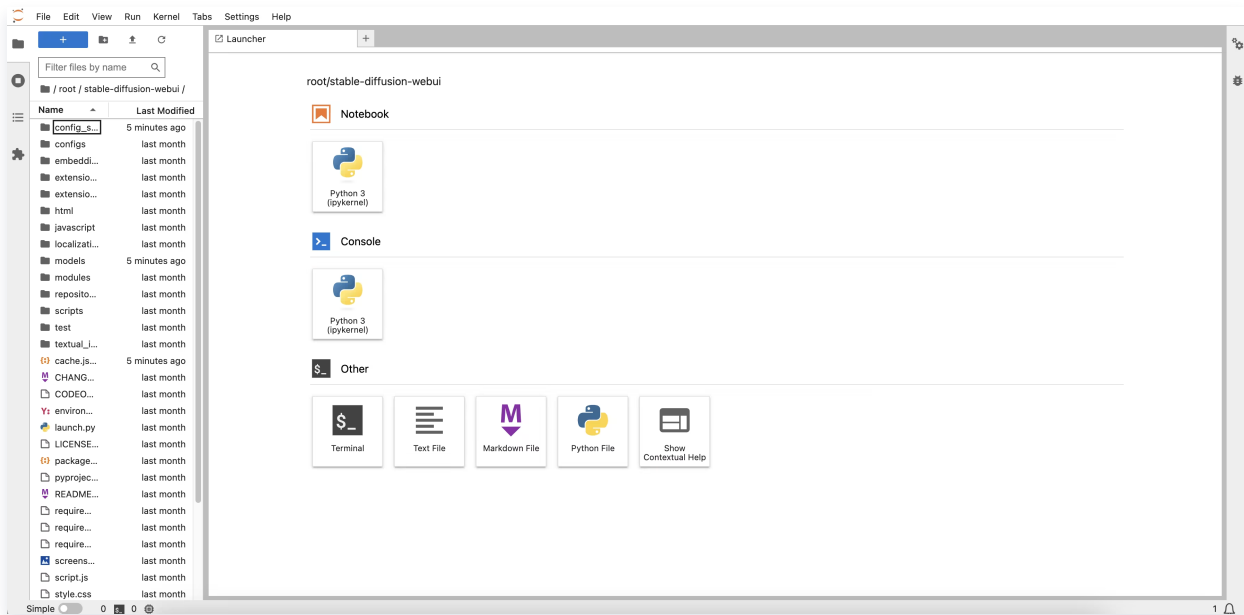
当您付费完成后，即完成了 Stable Diffusion 应用的创建。接下来，您可以登录实例并管理应用。

## 步骤3：连接实例

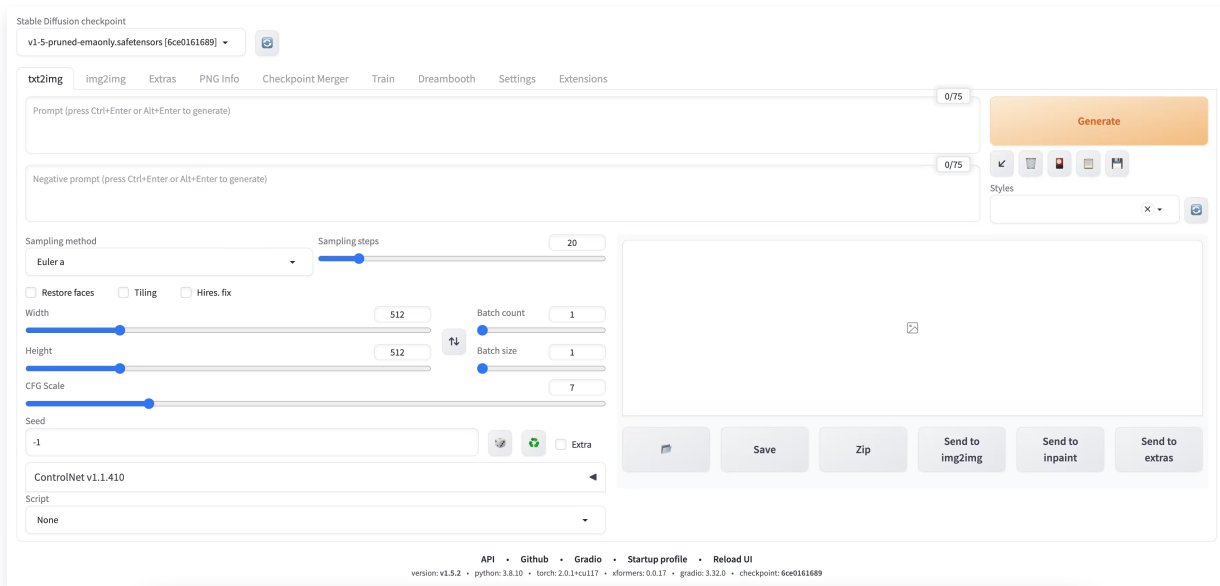
1. 登录 [高性能应用服务 HAI 控制台](#)，在服务器列表中，单击**算力连接**。进行以下两种登录方式。



○ 方式一：选择 JupyterLab 登录方式，进入 JupyterLab 终端。



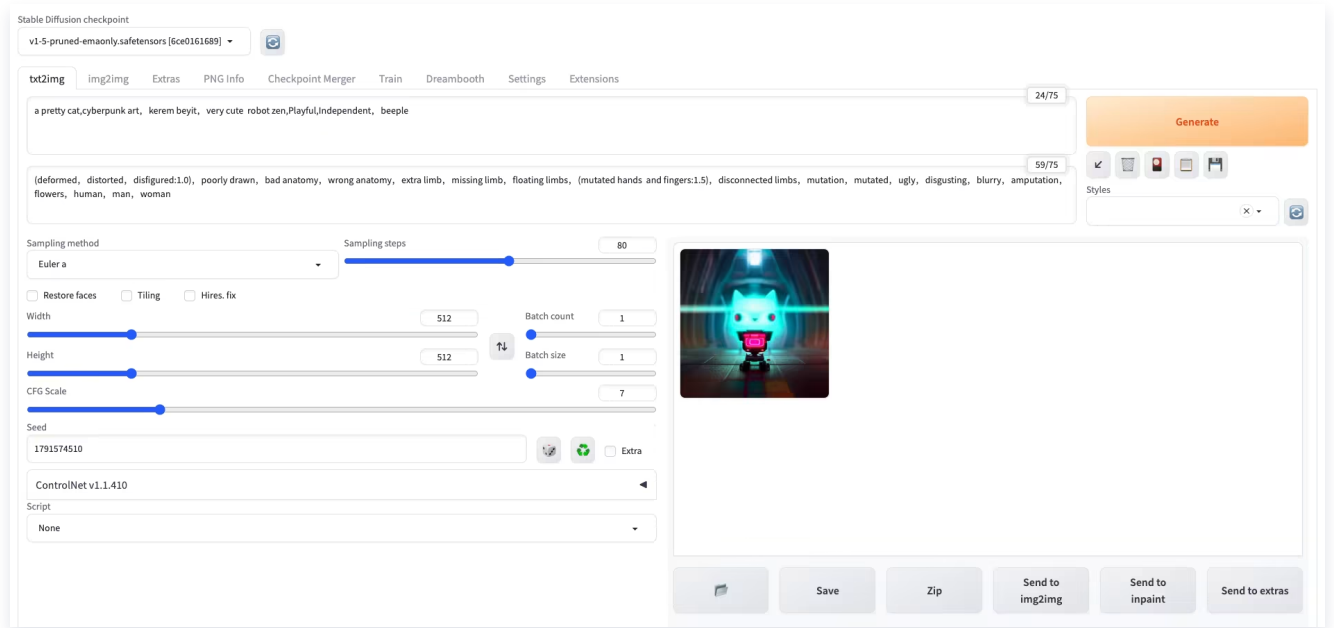
○ 方式二：选择 Gradio WebUI 登录方式，进入 Stable Diffusion checkpoint 可视化交互界面。



参考以下信息调整参数。

参数名	描述	值
Prompt	主要描述图像，包括内容风格等信息，原始的 WebUI 会对此处有字数的限制，您可以通过安装一些插件来突破字数的限制。	a pretty cat,cyberpunk art, kerem beyit, very cute robot zen,Playful,Independent, beepie
Negative prompt	为了提供给模型，您不需要的风格。	(deformed, distorted, disfigured:1.0), poorly drawn, bad anatomy, wrong anatomy, extra limb, missing limb, floating limbs, (mutated hands and fingers:1.5), disconnected limbs, mutation, mutated, ugly, disgusting, blurry, amputation, flowers, human, man, woman
CFG scale	分类器自由引导尺度，图像与提示符的一致程度。越低值产生的结果越有创意，数值越大成图越贴近描述文本。一般设置为7。	7
Sampling method	扩散算法的去噪声采样模式会影响最终效果，不同的采样模式的结果会有很大差异，一般是默认选择 euler。	Euler a
Sampling steps	在使用扩散模型生成图片时所进行的迭代步骤。需要注意的是，更高的迭代步数会消耗更多的计算时间和成本，但并不意味着一定会得到更好的结果。	80
Seed	随机数种子，生成每张图片时的随机种子。	1791574510

单击 **Generate** 即可生成图片，截图如下：



## 步骤4：销毁高性能应用服务 HAI（可选）

在使用结束后，在控制台算力管理页中，单击**更多 > 销毁**，即可销毁高性能应用服务 HAI，停止计费，结束使用。

