腾讯云

“中小企业”在线学堂

AI 2.0时代，如何通过AIGC打造爆款营销内容？

Stable Diffusion 基本原理介绍

导师介绍

腾讯云泛互联网
行业技术专家

刘远 johanyliu

10年+互联网从业经验，精通云原生，拥有丰富的微服务、大数据、机器学习开发和架构经验

目录 Menu

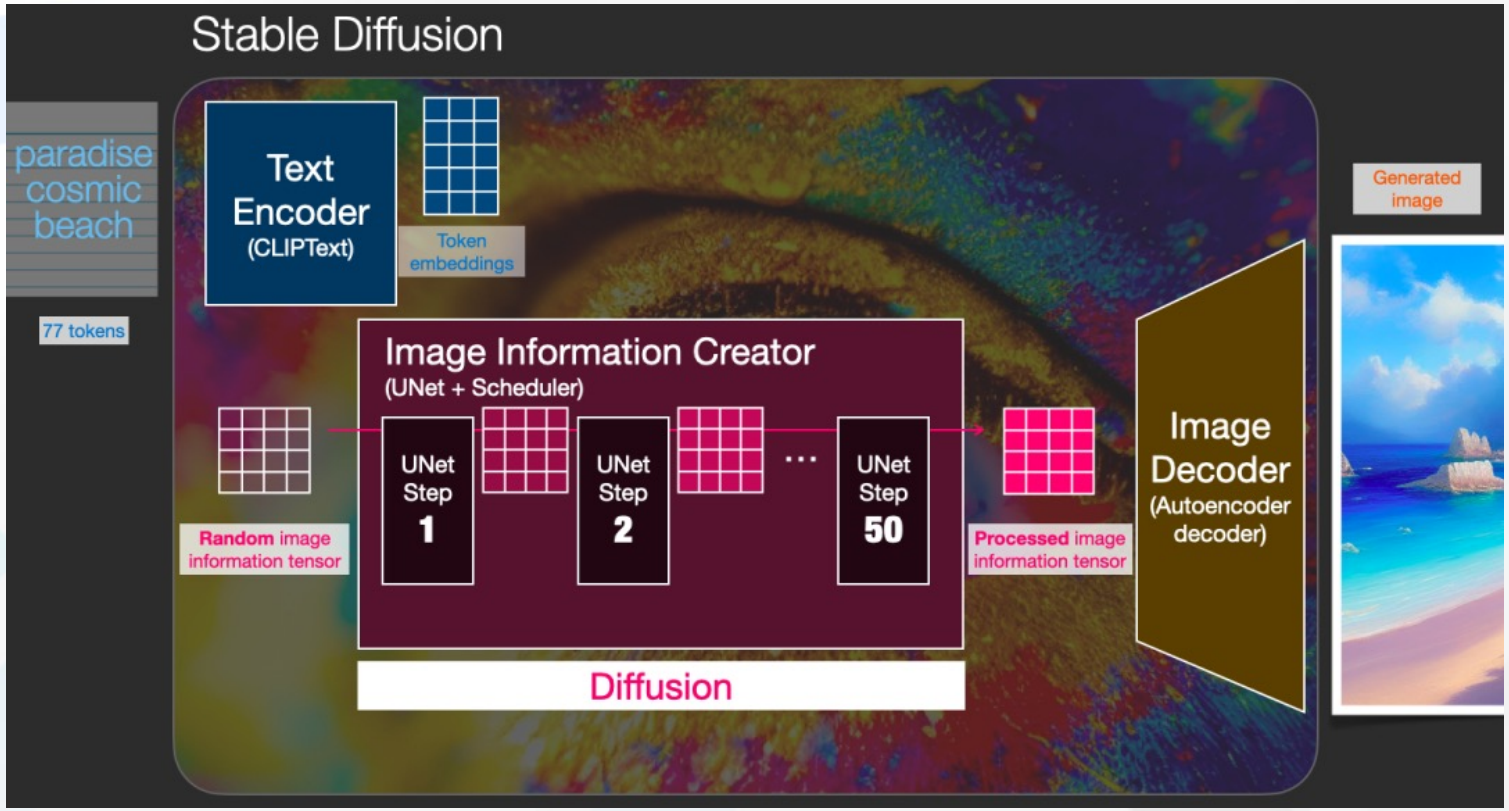
什么是 Stable Diffusion

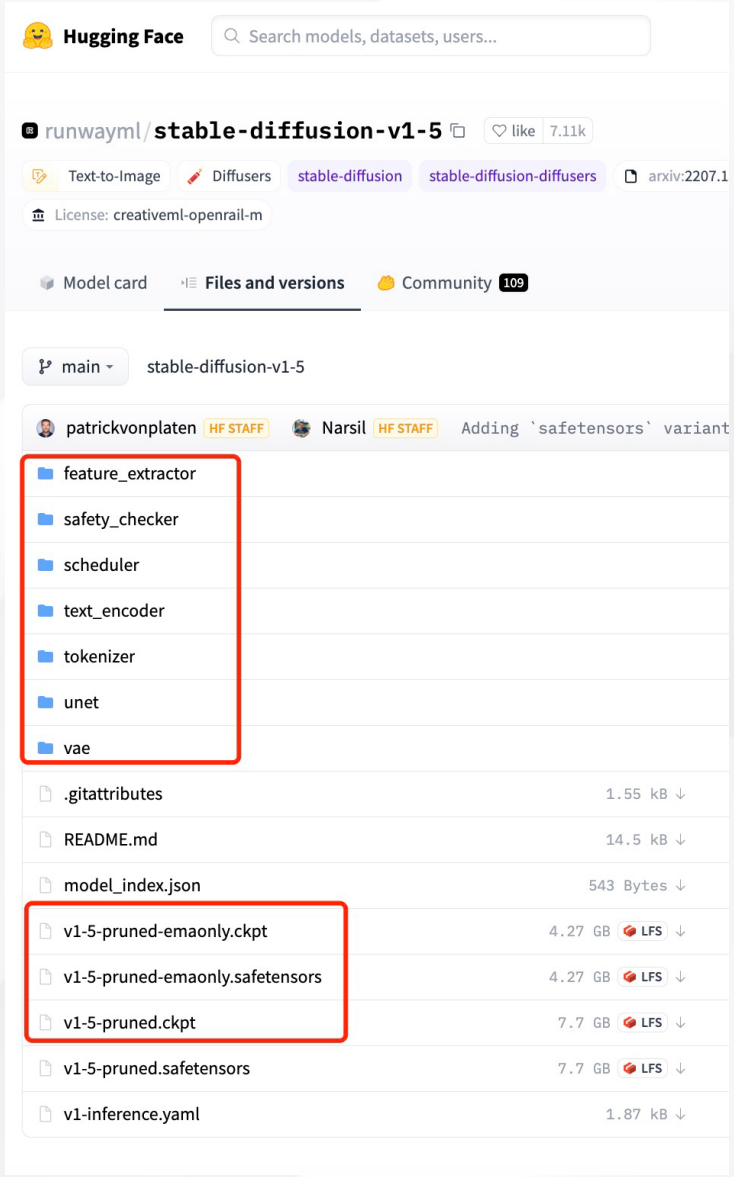
文生图：根据文本提示作为输入来生成的图像（输入为文本）

图生图：对图像根据文字描述进行修改（输入为文本+图像）

Stable Diffusion 系统架构图

- 每张图片都满足一定的规律分布，利用文本中包含的分布信息作为指导，将纯噪声的图片逐步去噪，生成一张跟文本信息匹配的图片
- Stable Diffusion 是一个组合系统，包含了多个模型子模块，组成扩散Pipeline



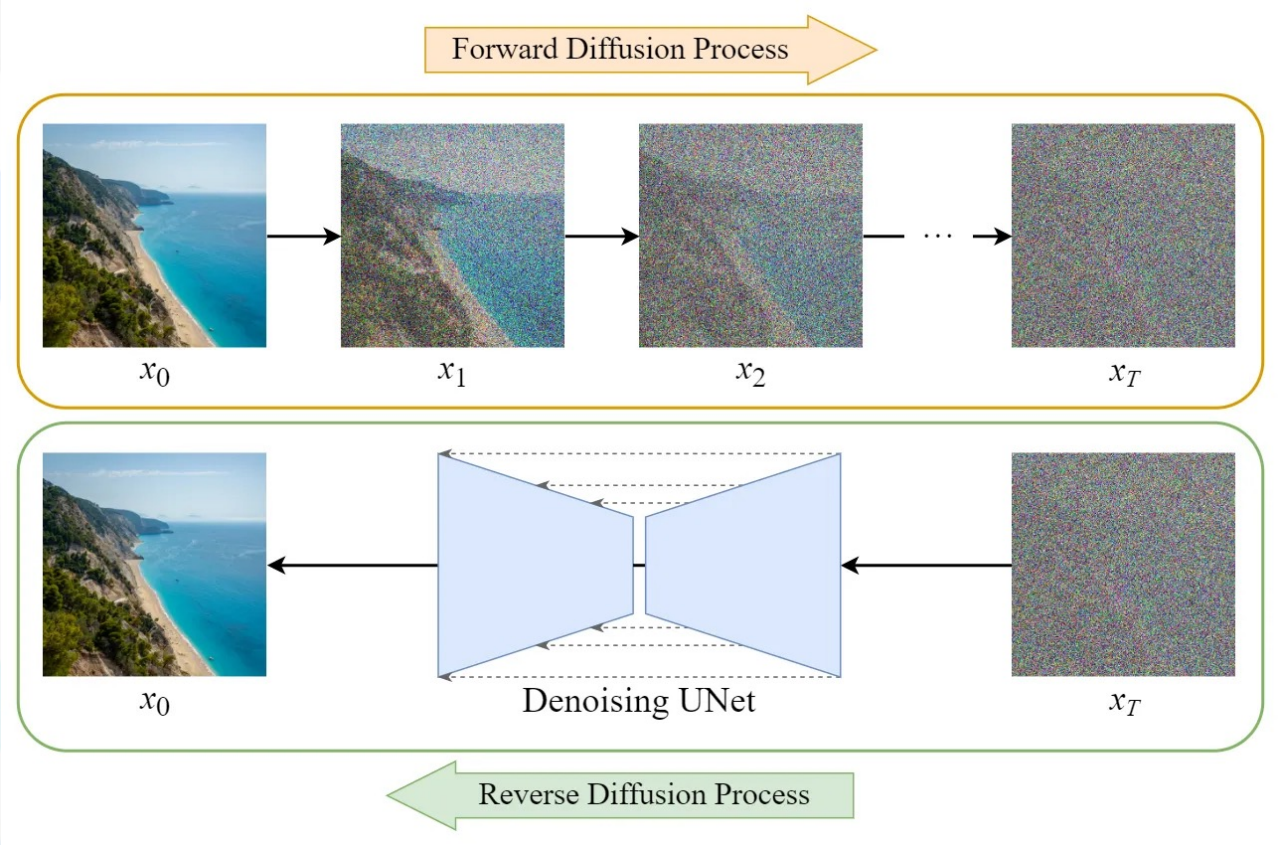


我们需要什么

- 一种生成图片的神经网络训练方法 —— **寻找分布的方法**
- 一种连接文字和图片的方法 —— **多模态连接的方法**
- 一种压缩和解压缩图片的方法 —— **加速训练的方法**

Diffusion model 运行机制

扩散模型——Diffusion

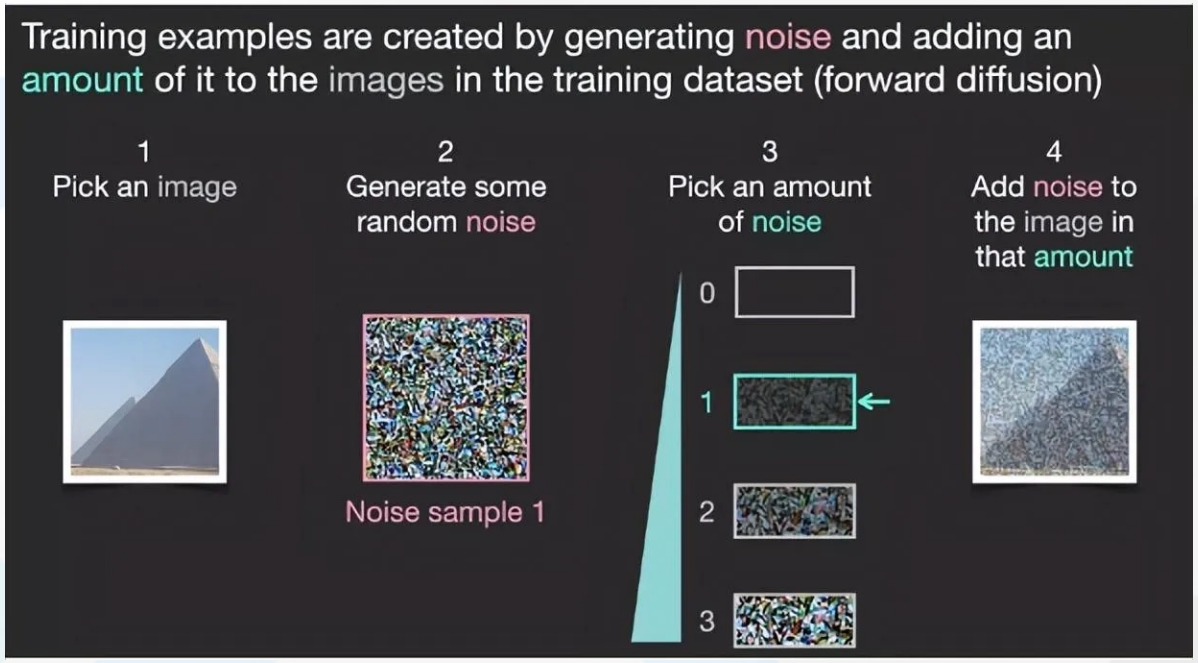


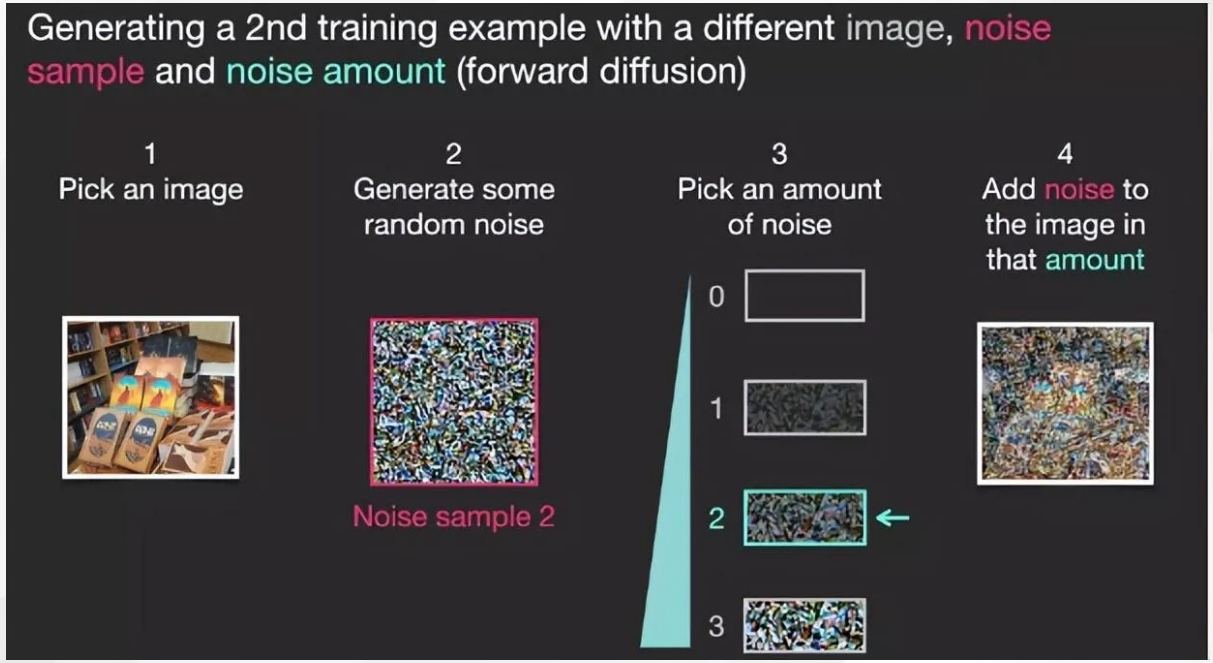
1. 寻找分布的方法

- 前向扩散过程（Forward Diffusion Process）向图片中逐渐引入噪声来破坏图像，直到图像变成完全随机的噪声
- 后向扩散过程（Backward Diffusion Process）使用一系列马尔可夫链逐步去除预测噪声，从高斯噪声中生成图片
- 扩散过程由 Unet 网络和 采样器（scheduler 算法）共同完成，在 Unet 网络中一步步执行生成过程， 其中采样器控制去噪方法和强度

数据集生成

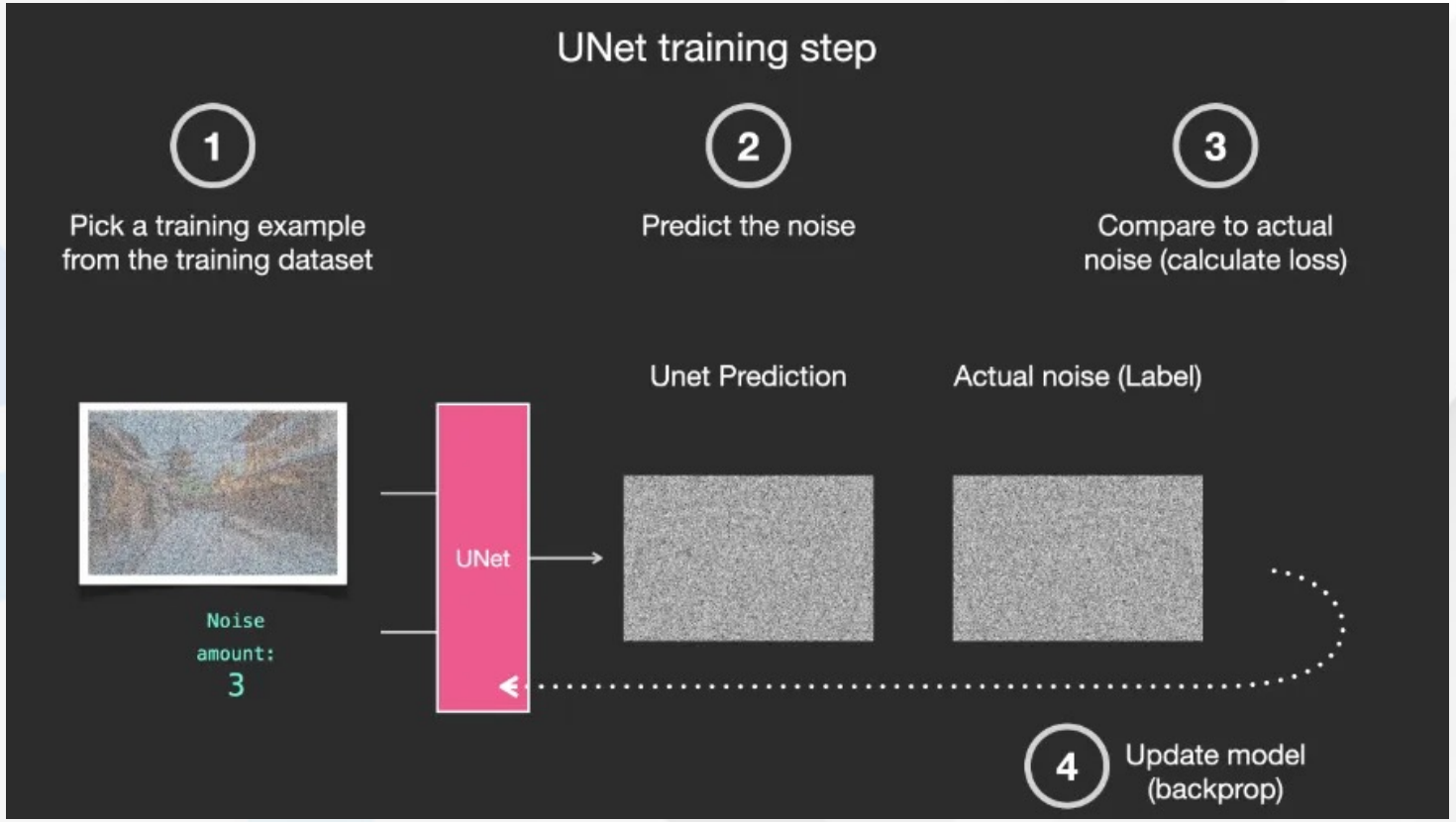
不断地向图片中添加不同强度的噪音（强度由 采样器 timestep 控制），为数据集中每一个图片都生成N个训练示例





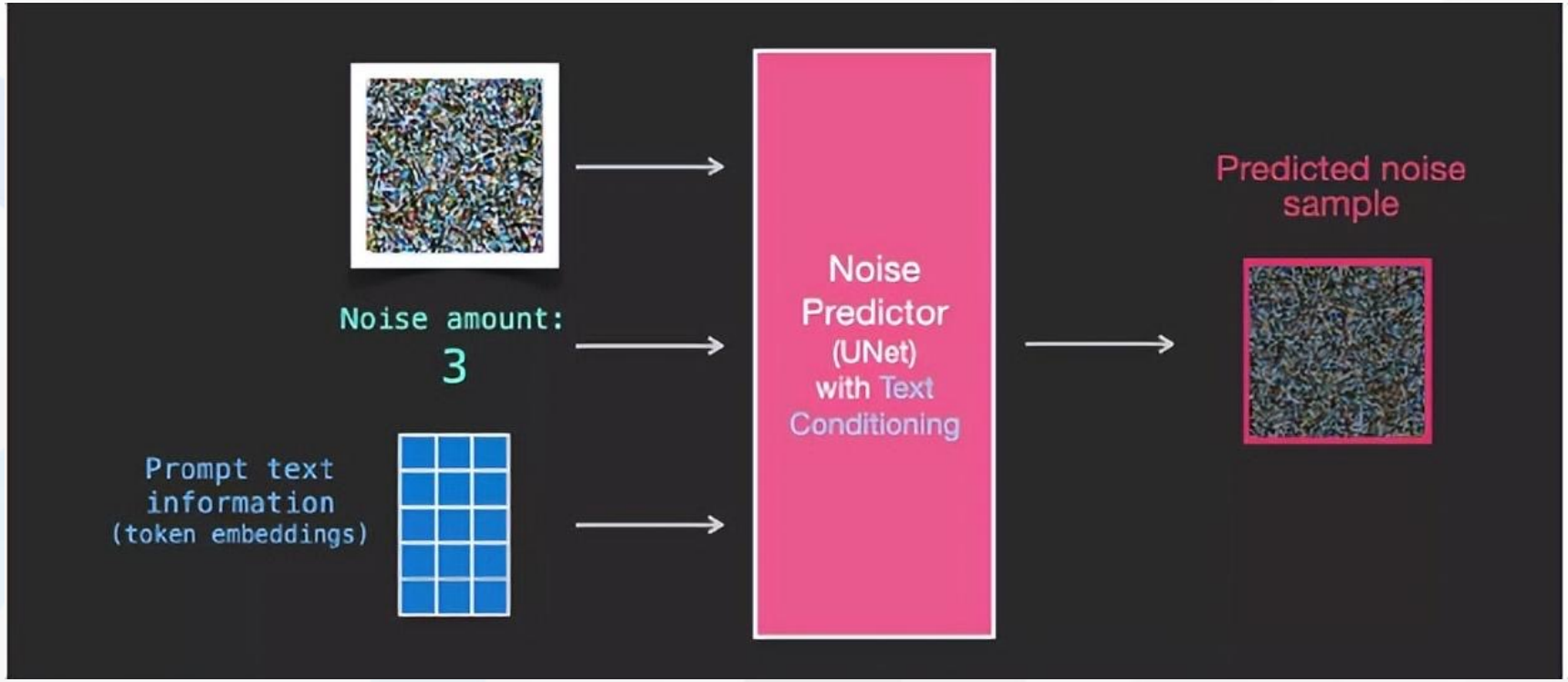
模型训练

使用上一步的数据集，对Unet模型进行训练，使用真实噪声标签作比较来反向传播更新，训练出“噪声预测器”



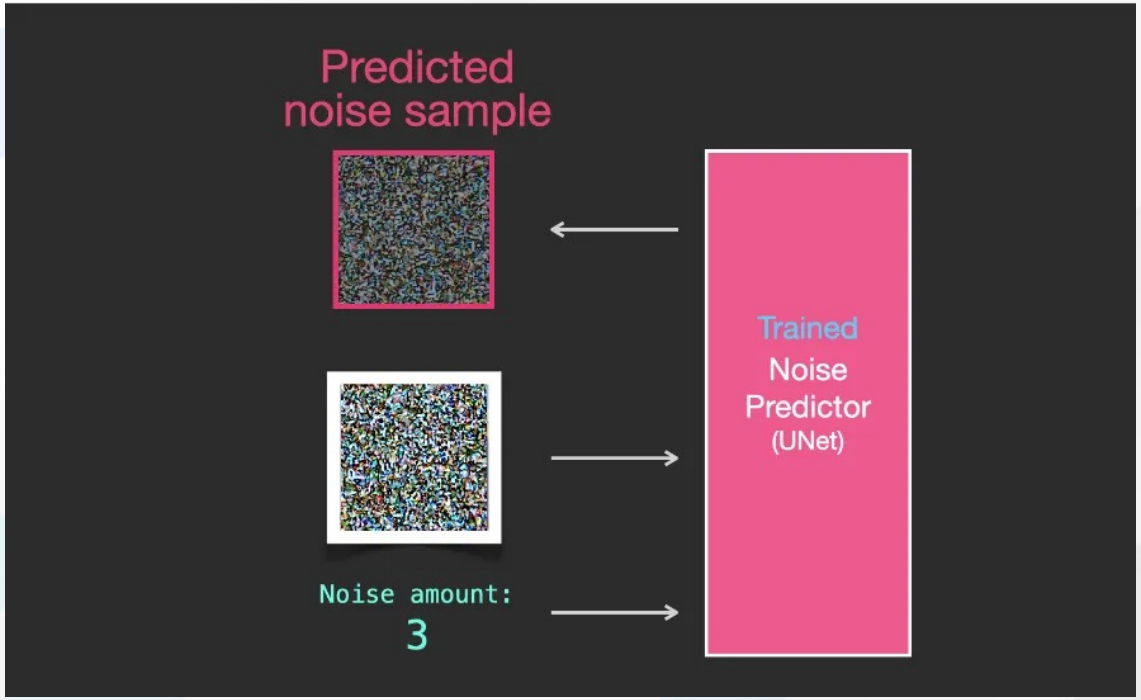
加入文本条件

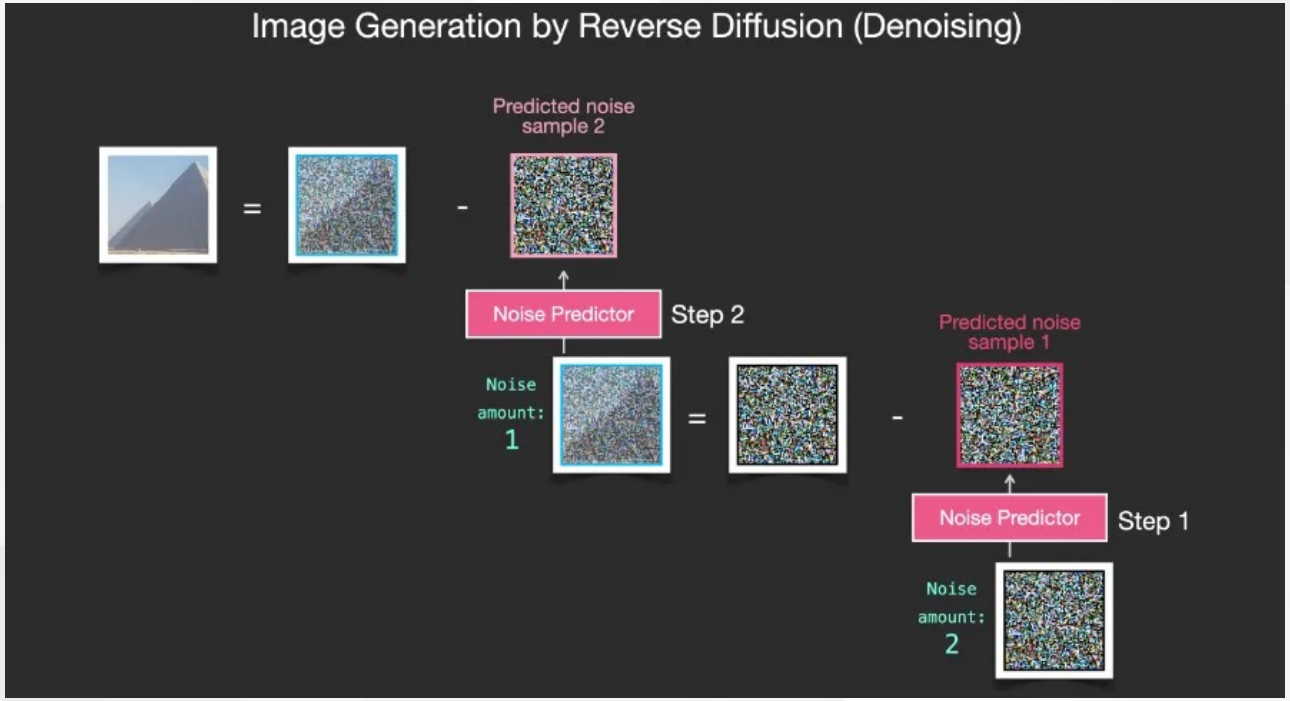
将文本信息作为附加控制向量，加入到模型向量中，通过注意力机制得到一个具有丰富语义信息的隐空间向量，引导图像往文本向量（Prompt）方向生成



降噪绘图

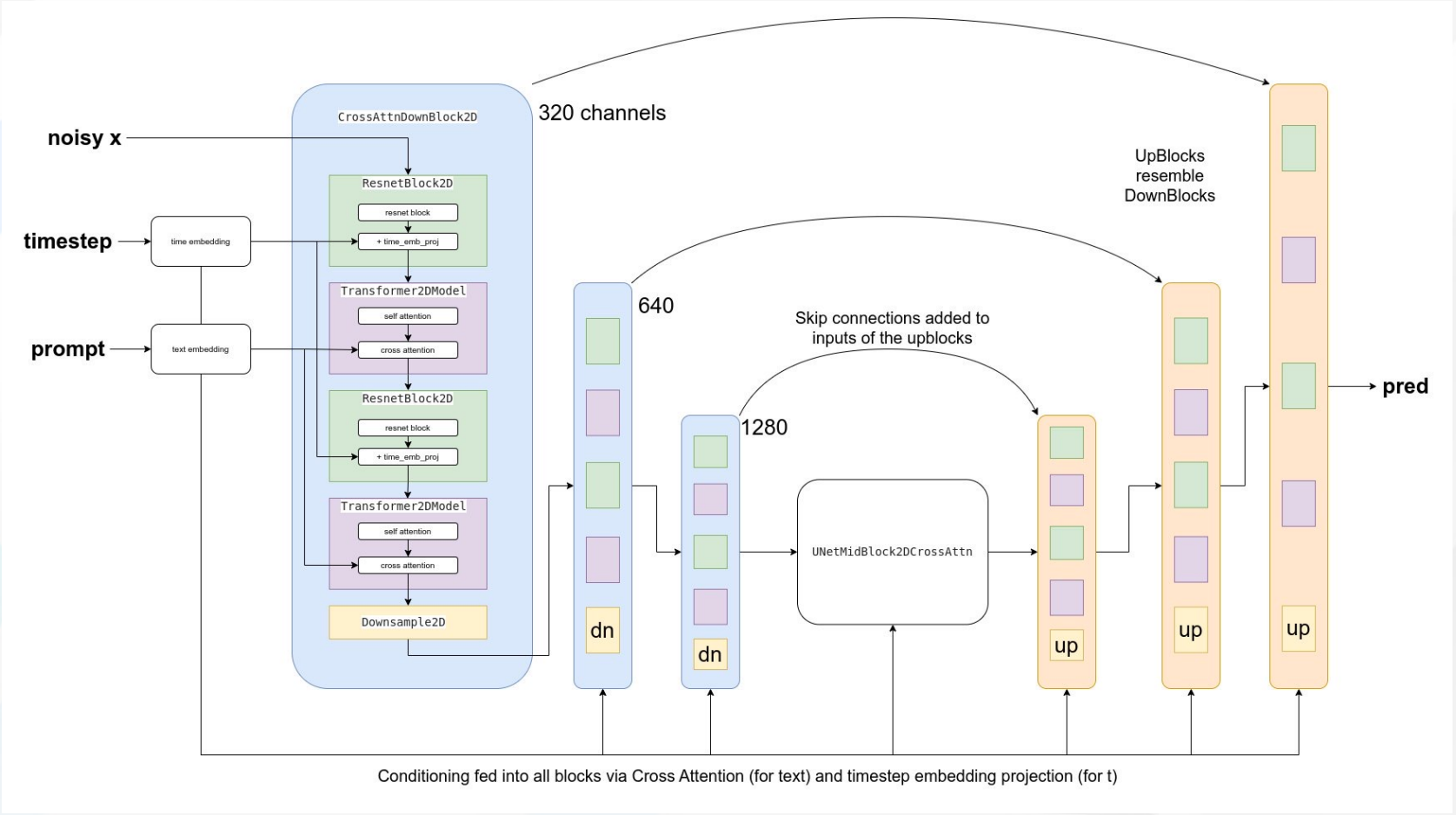
利用随机种子产出固定维度的噪声，使用训练好的Unet模型结合采样器，从每一步输出的样本中去掉噪音，得到具有文本信息的图像表征





Unet 网络

- Stable Diffusion 里的 Unet 模型采用 Encoder-Decoder 结构来预估噪声
- 使用 DownSample 和 UpSample 进行样本的下上采样，采样模块之间的 ResBlock 和 SpatialTransformer，分别接收 timesteps （采样器使用）和文本信息作为输入
- ResBlock不直接处理文本向量，由SpatialTransformer混合进潜空间向量，在下一层得以应用



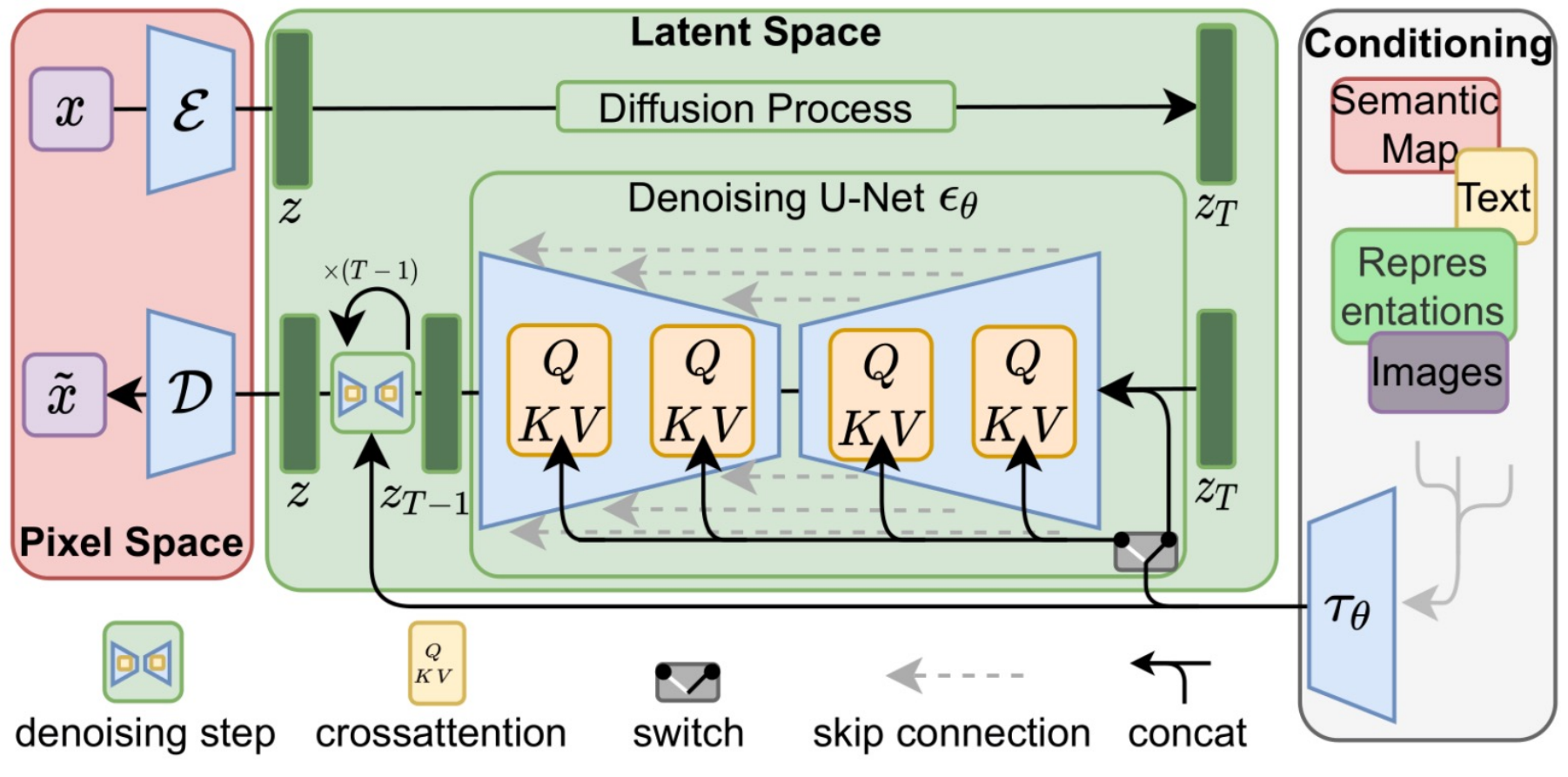
Unet 模型输入包括三部分

a. 图像潜空间向量【batchsize，隐空间通道，图片高度/8，图片宽度/8】
b. timesteps【batchsize】
c. 文本信息向量【batchsize，文本最大编码长度，向量大小】

Stable Diffusion 模块分析

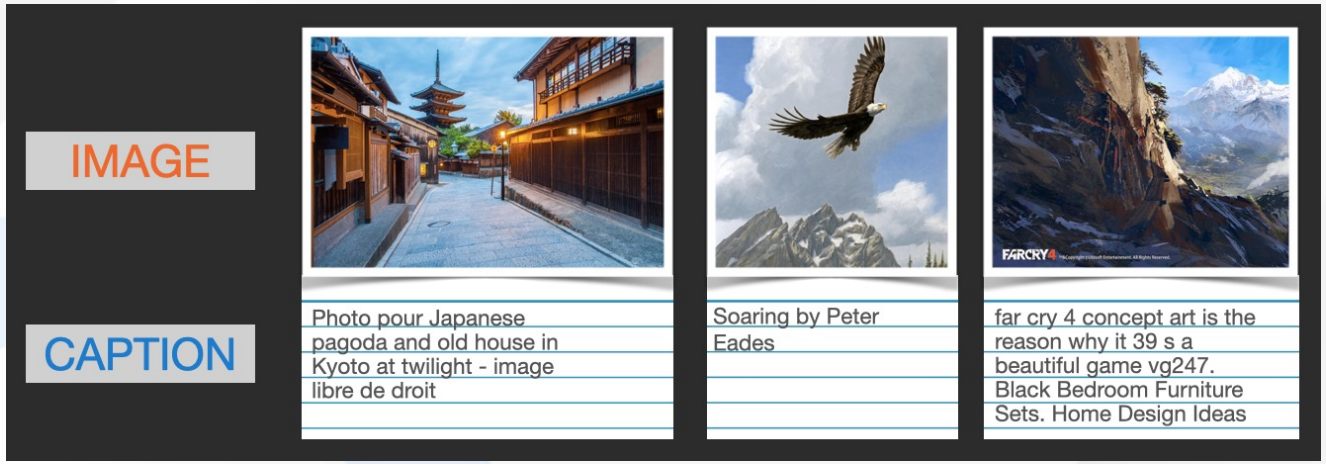
Stable Diffusion

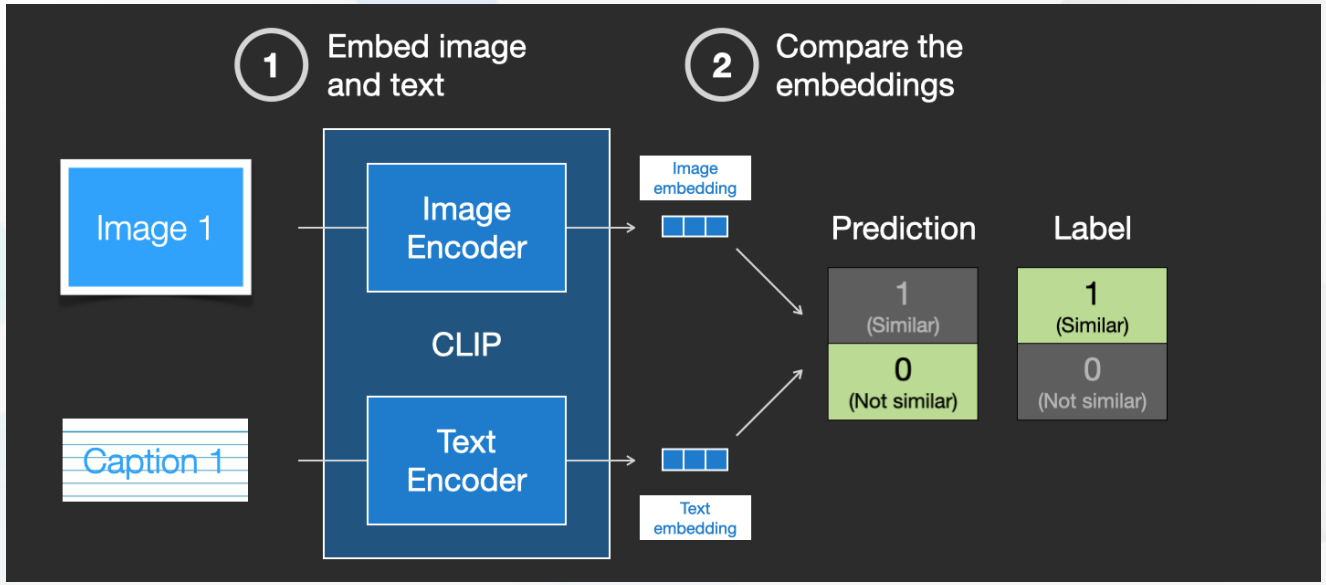
- 为Unet模型添加条件控制机制（text、images、inpainting）
- 在潜空间进行diffusion过程，而不是像素空间



CLIP 模型

2. 多模态连接的方法





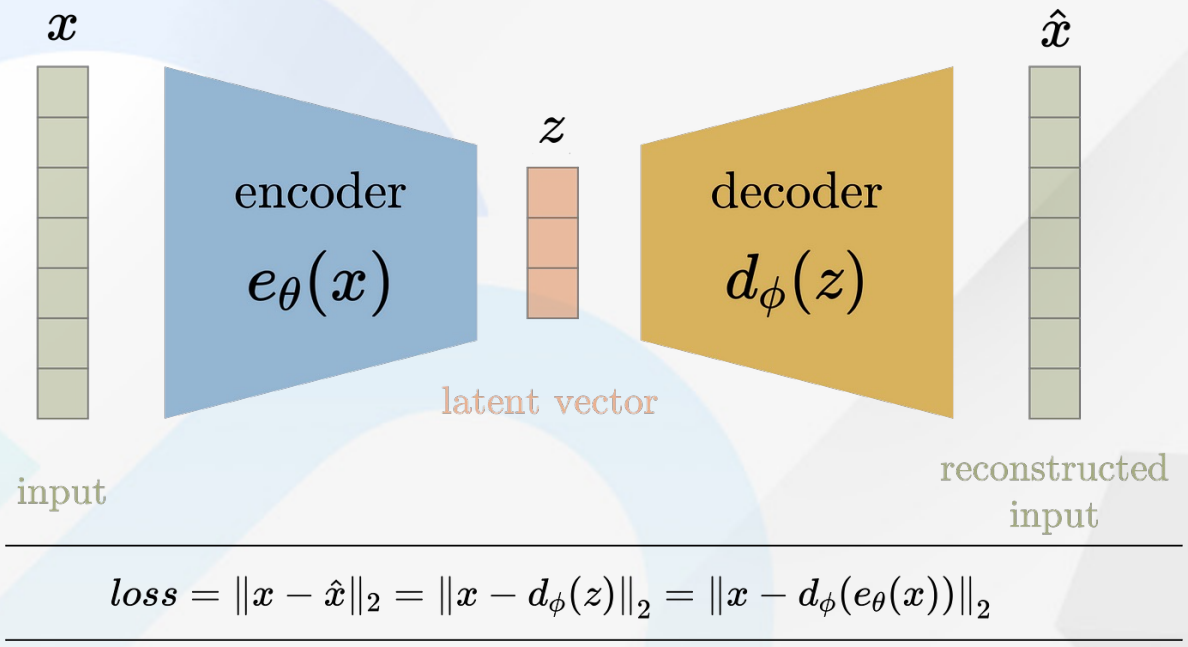
- CLIP是图像 Encoder 和文本 Encoder 的组合，CLIP 训练时分别对图像和文本进行编码，使用余弦相似度进行对比，反向更新两个 Encoder 的参数
- 完成训练后，输入配对的图片和文字，两个 encoder 就可以输出相似的 embedding 向量，输入不匹配的图片和文字，两个 encoder 输出向量的余弦相似度就会接近于 0
- 推理时，输入文字通过 CLIP 转化为 embedding，映射进 Unet 注意力层，和图片相互作用

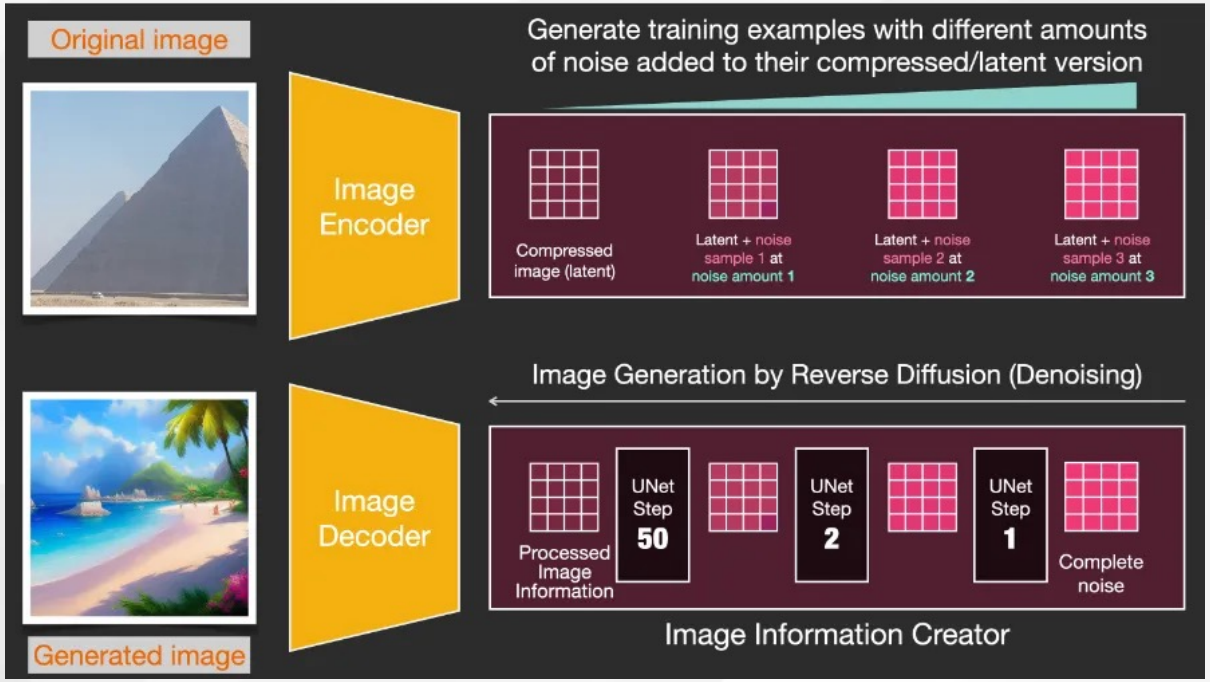
变分自编码器 VAE

3. 加速训练的方法

为了加快图像生成过程，Stable Diffusion 没有在像素图像上进行运行，而是在图像的压缩版本（潜空间）上运行。

模型学习时，将图像的尺寸先减小再恢复到原始尺寸。使用解码器从压缩数据中重建图像时，会同时学习之前的所有相关信息。

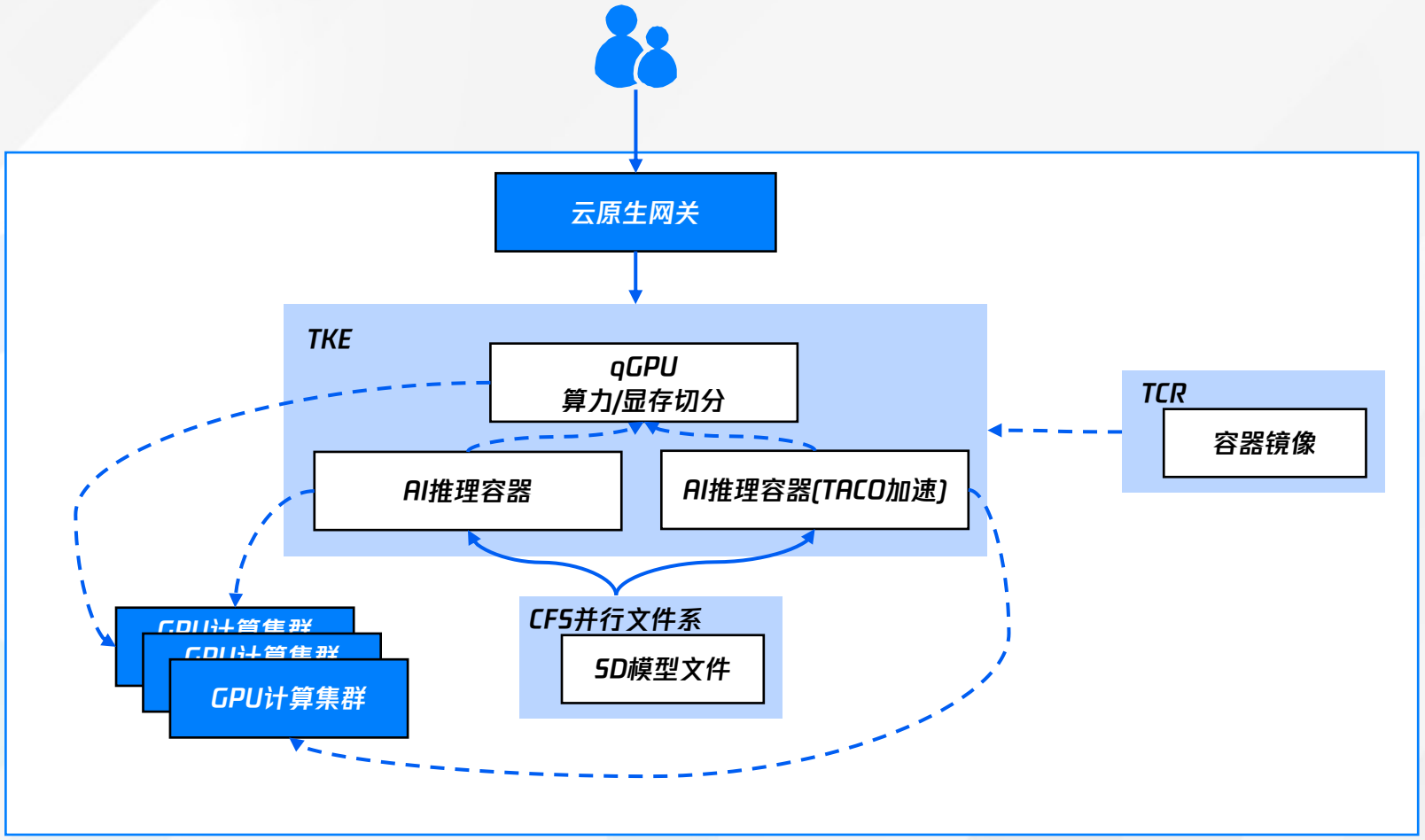




感谢观看！
Thank you

Stable Diffusion 腾讯云云原生容器部署实践

分享大纲



1.使用腾讯云TKE+CFS部署Stable Diffusion

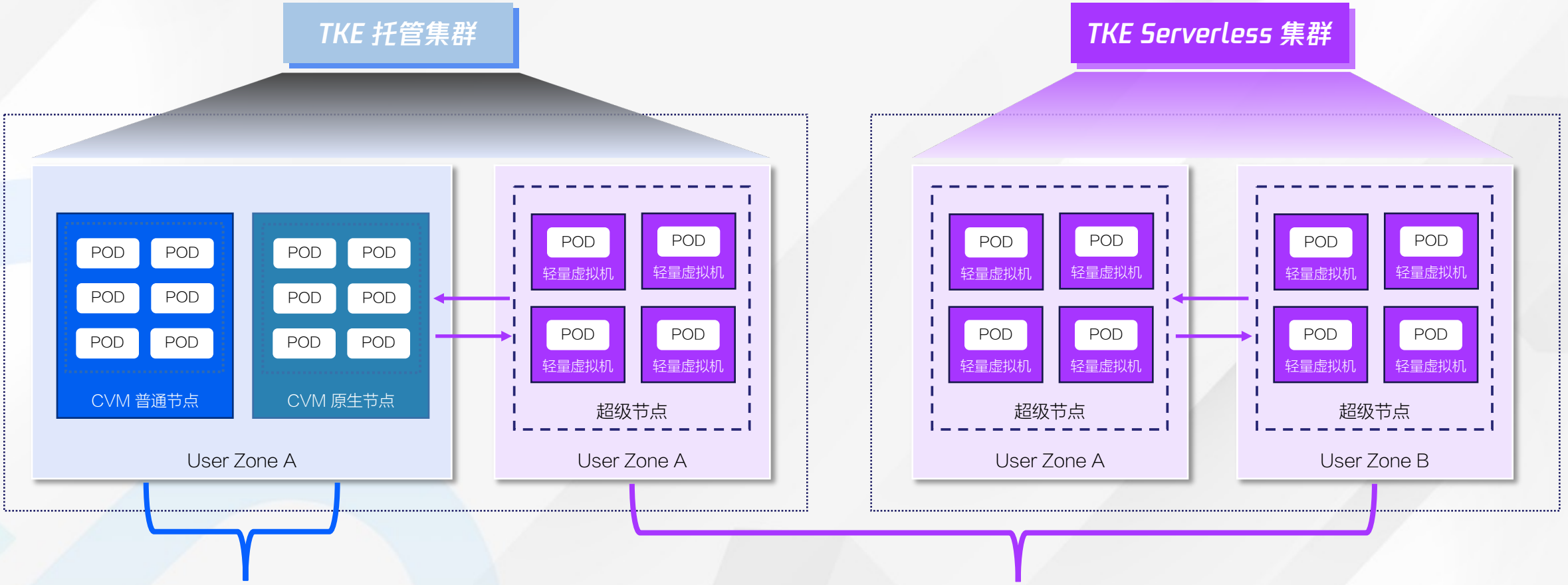
2.通过腾讯云云原生 API 网关对外提供 SD服务

3.通过qGPU提升推理服务并发性能

4.通过TACO优化推理速度

使用腾讯云TKE+CFS部署Stable Diffusion

TKE：基于原生 K8S提供以容器为核心的解决方案



- 按照集群资源付费
- 需要运维集群节点
- 容器共享OS内核
- 安全隔离性弱

- 按照Pod资源付费
- 无需运维集群节点、极致弹性
- 容器独立内核
- 虚拟机级别隔离性

容器存储方案

块存储 CBS-CSI

- 支持 TKE 集群通过控制台快捷选择存储类型，并创建对应块存储云硬盘类型的 PV 和 PVC。
- 支持拓扑感知、在线扩容、快照和恢复。
- 适合 POD 单挂，低延迟,高 IOPS 场景。

网络文件系统 CFS-CSI

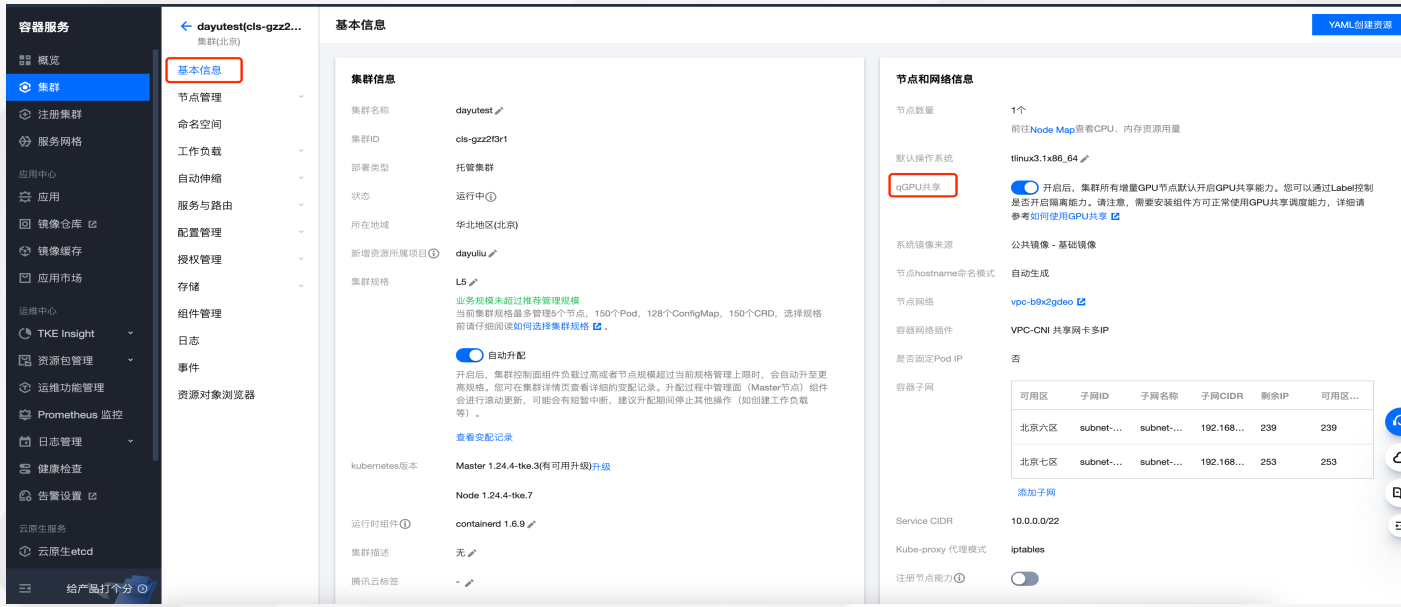
- 通过 CFS-CSI 扩展组件，您可以快速在容器集群中通过标准原生 Kubernetes 使用 CFS。
- 支持标准存储和性能存储。
- 适合 POD 多挂，CVM 和 POD 共享文件，无需扩容。

对象存储 COS-CSI

- 通过 COS-CSI 扩展组件，您可以快速的在容器集群中通过标准原生 Kubernetes 以 COSFS 的形式使用 COS。
- 适合程序使用文件系统访问 cos 里面的文件。

准备待部署 SD的 TKE 集群

1. 开通并创建 TKE 集群
2. 集群选择托管类型，Worker节点选择【GPU计算型PNV4】- A10，安装 GPU470 驱动，CUDA 版本11.4.3，cuDNN 版本 8.2.4，见下图。
3. 根据部署对 GPU 共享的需求，可选择开启 qGPU。



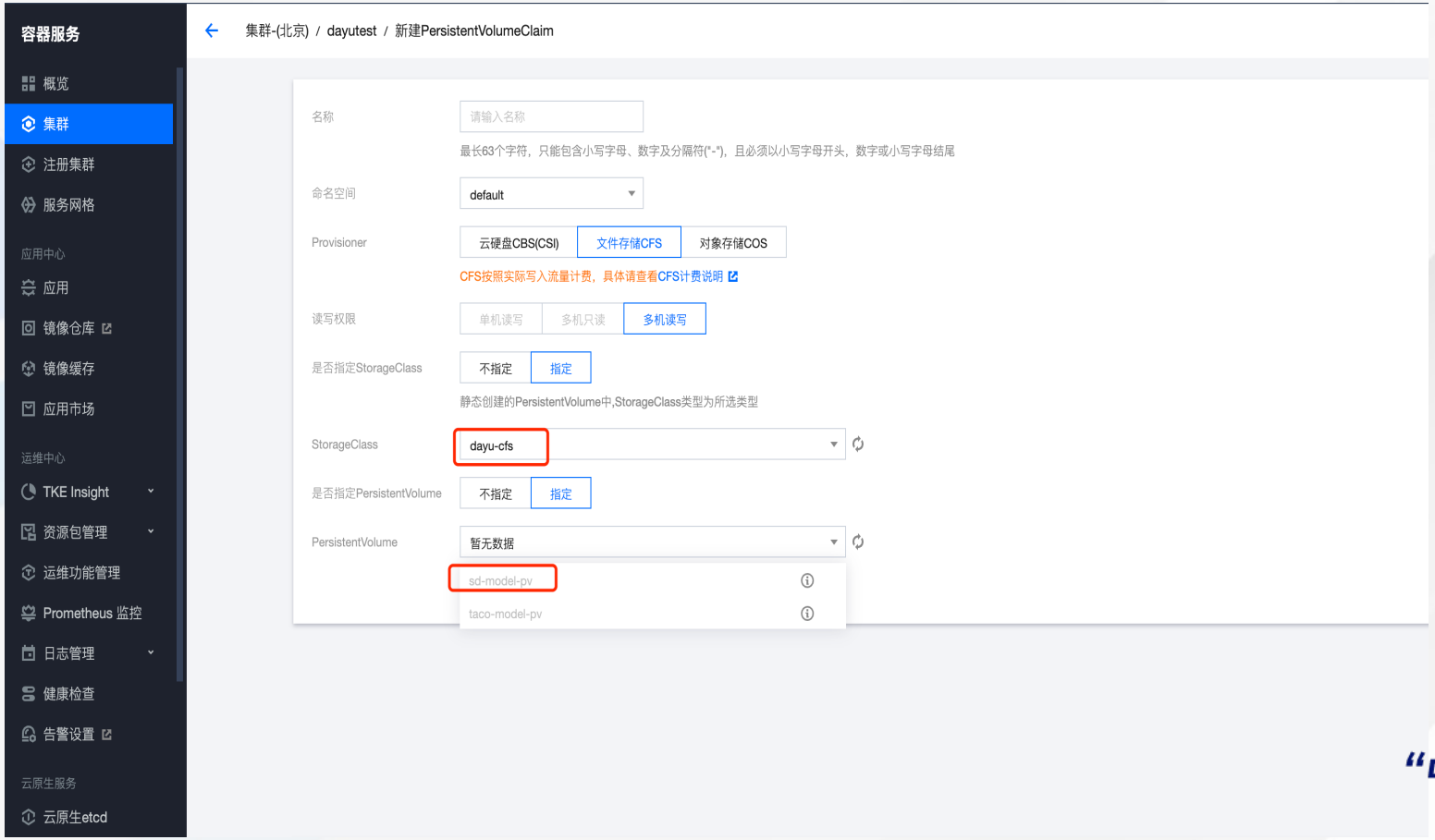
使用CFS 保存SD模型文件

1. 创建存放模型的文件存储 CFS
2. 选择与集群相同的 VPC 和子网，开通 CFS 服务。
3. 在 CFS 远程挂载点，新建 /models/Stable-diffusion 目录。下载 v1-5-pruned-emaonly.safetensors 模型文件至/models/Stable-diffusion


```
[root@VM-128-5-tencentos Stable-diffusion]# pwd
/taco/models/Stable-diffusion
[root@VM-128-5-tencentos Stable-diffusion]# ls
v1-5-pruned-emaonly.safetensors
[root@VM-128-5-tencentos Stable-diffusion]#
```

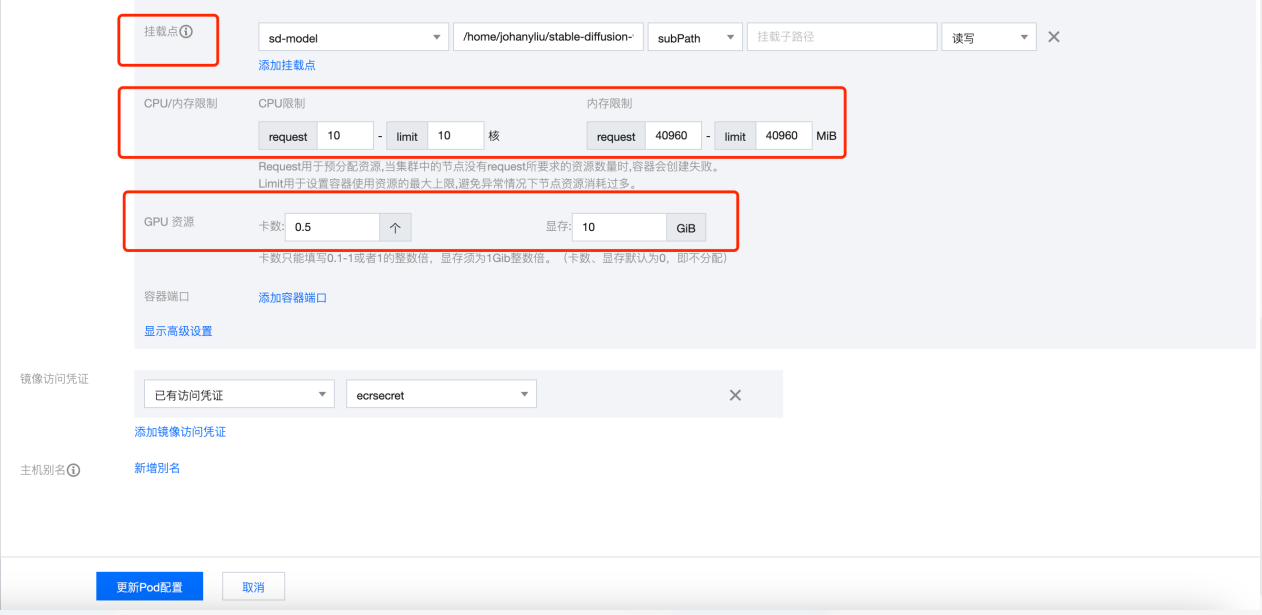
通过TKE挂载CFS

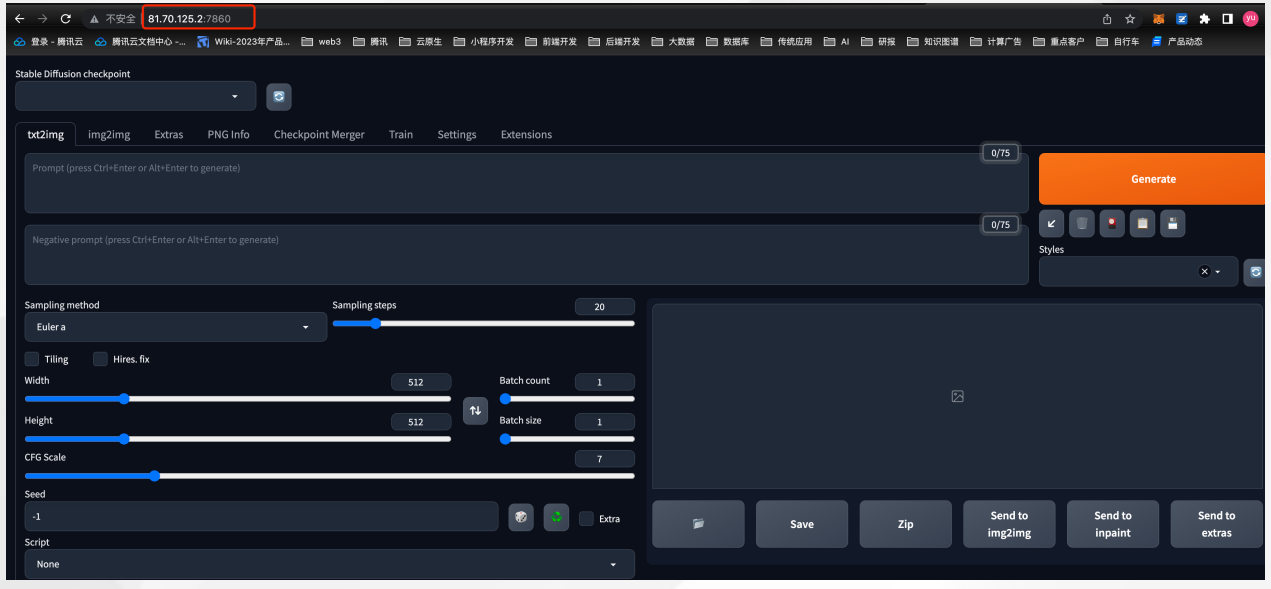
1. 在 TKE 控制台上，创建 CFS 类型 StorageClass，选择共享实例
2. 使用CFS里新建的/models/Stable-diffusion目录，以及上面的StorageClass，静态创建PV&PVC



创建 SD Web UI 工作负载

1. 在TKE控制台上，选择【工作负载】-【Deployment】-【新建】，部署SD-webui 镜像
2. 在部署页面里，【数据卷】选择【添加数据卷】。【使用已有PVC】方式，添加前面创建的 PVC
3. 【实例内容器】-【镜像】部分，选择已保存在 TCR 里的 stable-diffusion-webui 镜像
4. 将 GPU 资源的卡数设置为 1，如果开启了 qGPU，这里还可以填写0.1-1之间的数值
5. 创建 Deployment 对应的 Service，选择【公网LB访问】，对外暴露7860端口访问
6. 通过 CLB 公网 IP 地址，访问SDWeb UI 服务。如果有限流或访问控制需求，推荐选择云原生网关





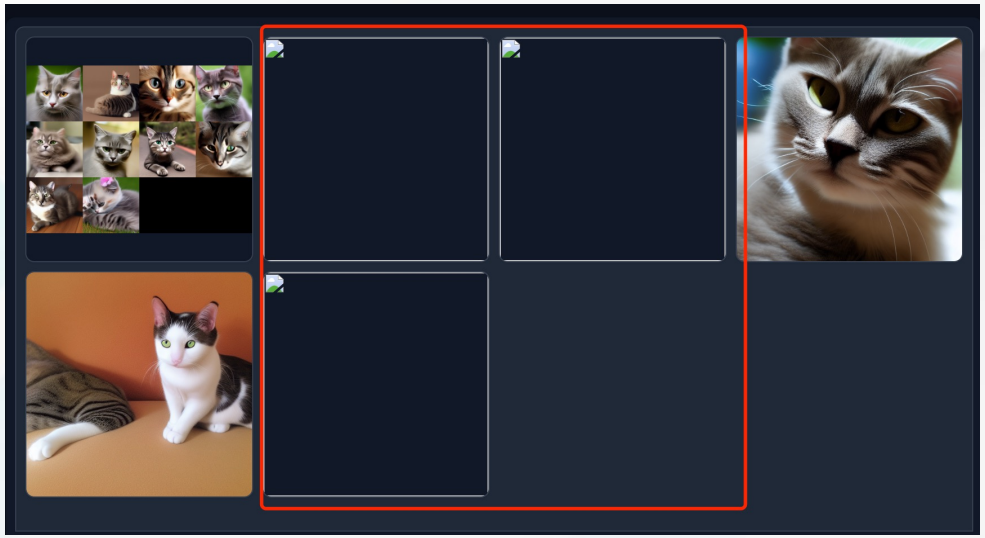
通过腾讯云云原生 API 网关对外提供 SD服务

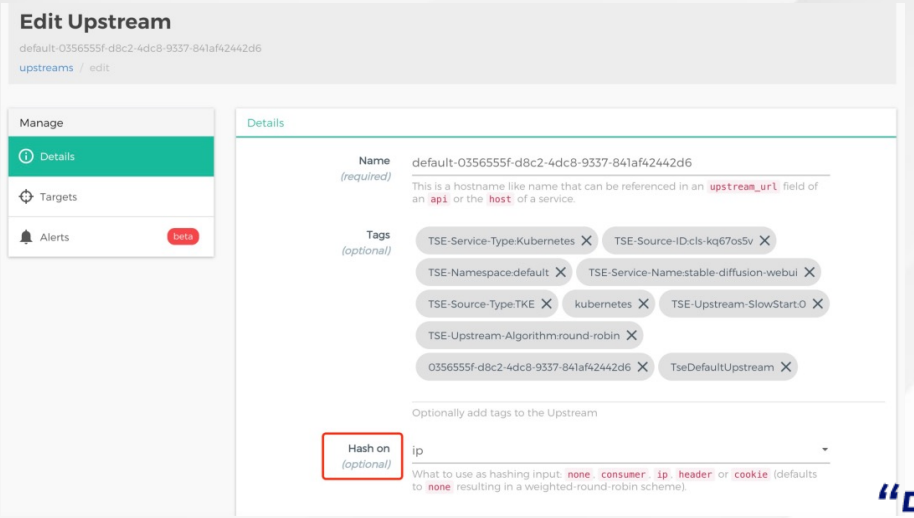
云原生API网关：云上微服务架构的流量入口

云原生API网关（Cloud-Native APIGateway）是腾讯云基于开源网关推出的一款高性能高可用的API生命周期管理产品，100%完美兼容开源网关Kong的API，减少用户自建网关的开发及运维成本。

通过云原生 API 网关对外提供 SD服务

1. 开通云原生网关，选择和 TKE 集群、CFS 同 VPC 的实例
2. 在网关控制台上，选择【路由管理】-【服务来源】，绑定TKE集群
3. 选择【路由管理】-【服务】，新建网关服务。在【服务列表】中，选择部署Deployment 时启用的 Service 进行映射
4. 点击服务名，新建访问路由。将请求方法设置为【ANY】，Host 填写云原生网关的公网 IP
5. Stable Diffusion Web UI 出图时会进行多轮请求，将 Deployment 的 Pod 副本数量修改为大于1时，还需要配置 Session 会话保持，保证同一IP的客户请求落在相同的Pod里。选择【路由管理】-【Konga控制台】，找到Konga公网访问地址，在Konga控制台里找到【UPSTREAM】，点击【DETAILS】，在【HASH ON】下拉框里，选择【IP】，完成基于客户端IP的会话保持配置



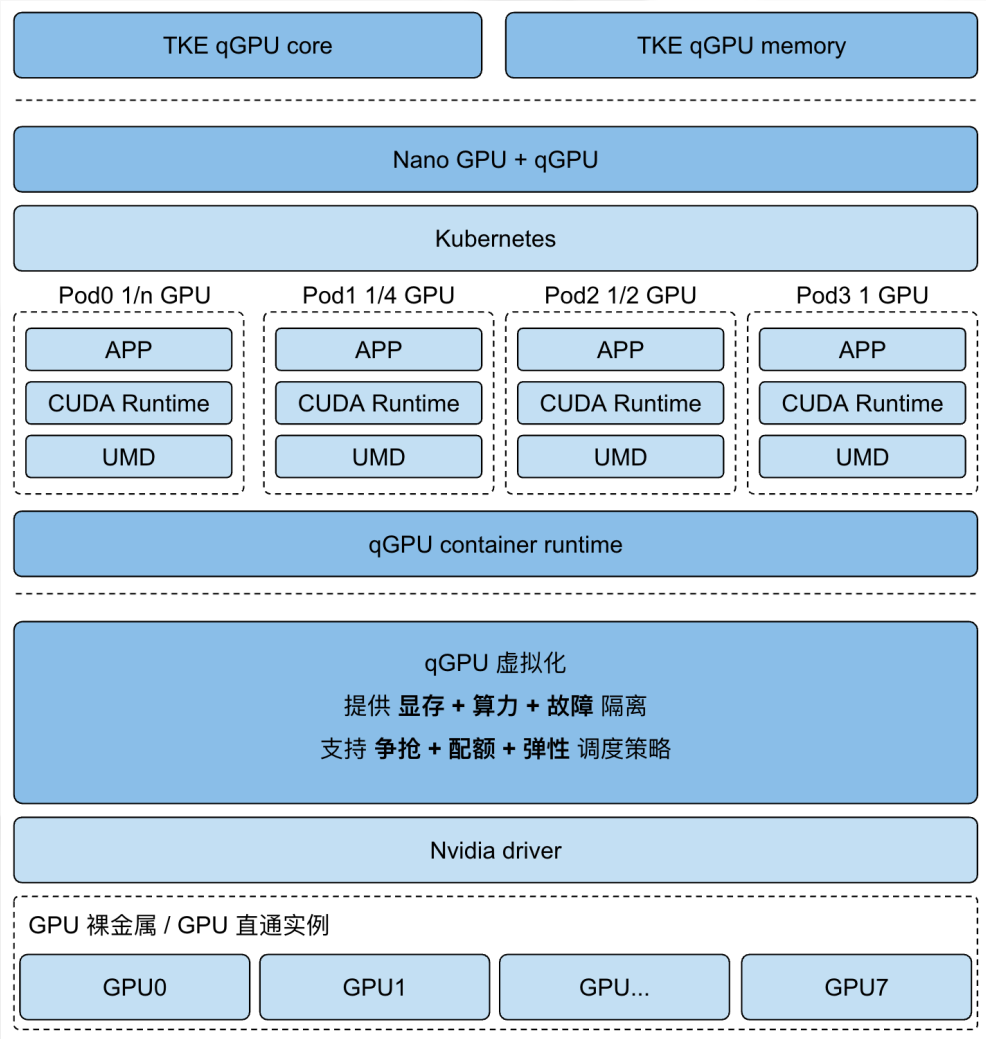


通过qGPU提升推理服务并发性能

qGPU：GPU共享技术

qGPU 是腾讯云推出的 GPU 共享技术，支持在多个容器间共享 GPU 卡并提供容器间显存、算力强隔离的能力，从而在更小粒度的使用 GPU 卡的基础上，保证业务安全，达到提高 GPU 使用率、降低客户成本的目的。

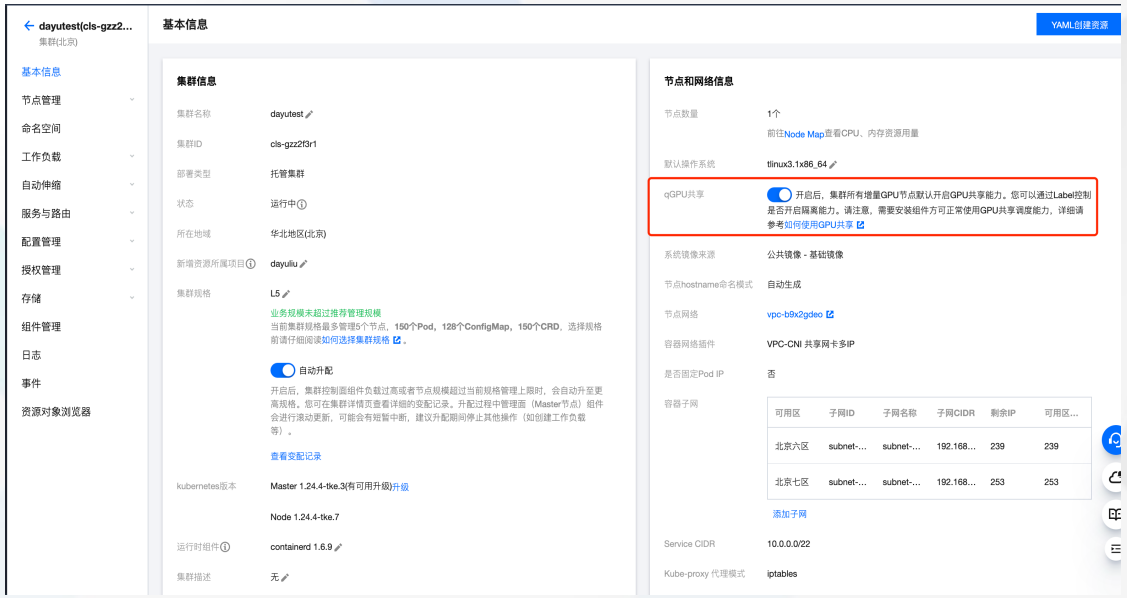
- 灵活性：自由配置 GPU 显存大小和算力占比
- 云原生：支持标准 Kubernetes 和 NVIDIA Docker
- 兼容性：业务不重编、CUDA 库不替换、业务无感
- 高性能：GPU 设备底层虚拟化，高效收敛，吞吐接近0损耗
- 强隔离：支持显存和算力的严格隔离



Stable Diffusion 使用 qGPU

Stable Diffusion Web UI 服务以串行方式处理请求，如果希望增加推理服务并发性能，可以考虑扩展Deployment的Pod数量，以轮询的方式响应请求。这里我们采用TKE qGPU能力，将多个实例 Pod 运行在同一张 A10 卡上，在保障业务稳定性的前提下，切分显卡资源，降低部署成本。

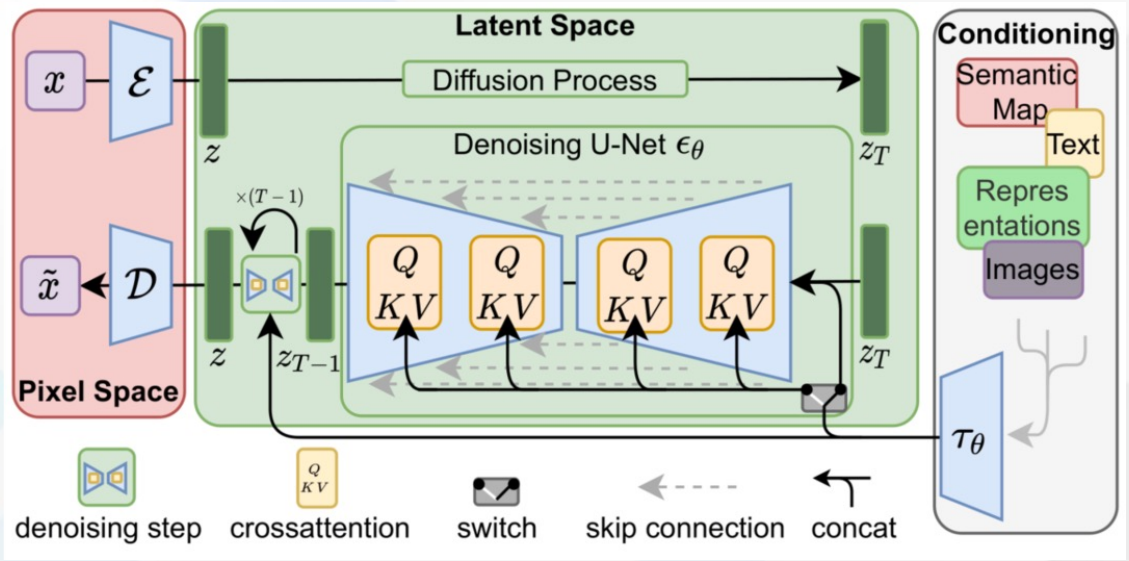
采用 qGPU 方式，需要先将 Pod 的资源申请方式进行修改。如果计划单卡上部署2个 Pod，将卡数改为50%的算力分配，显存设置为 A10 显存的一半。Deployment YAML更新完成后，调整 Pod 数量为2个，即可实现负载均衡的 Stable Diffusion 轮询模式



通过TACO优化推理速度

TACO: 优化 Stable Diffusion 推理性能

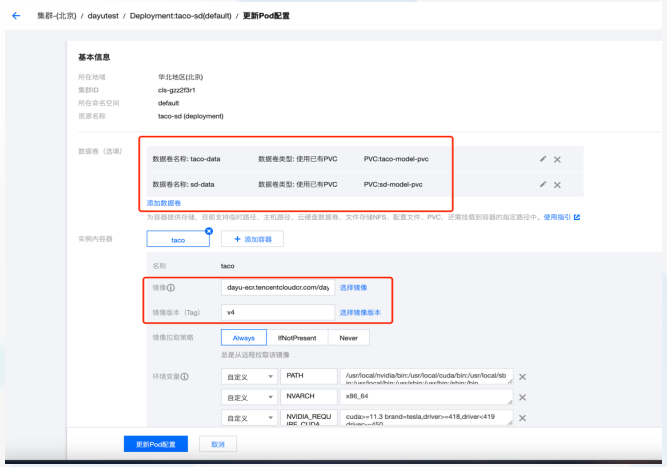
腾讯云计算加速套件 TACO Kit（TencentCloud Accelerated Computing Optimization Kit）是一种异构计算加速软件服务，具备领先的 GPU 共享技术和业界唯一的 GPU 在离线混部能力，搭配腾讯自研的软硬件协同优化组件和硬件厂商特有优化方案，支持物理机、云服务器、容器等产品的计算加速、图形渲染、视频转码各个应用场景，帮助用户实现全方位全场景的降本增效。在 Stable Diffuison 场景中，可以实现模型的前向推理能力提升，出图由原来的约 2 秒缩短到 1 秒

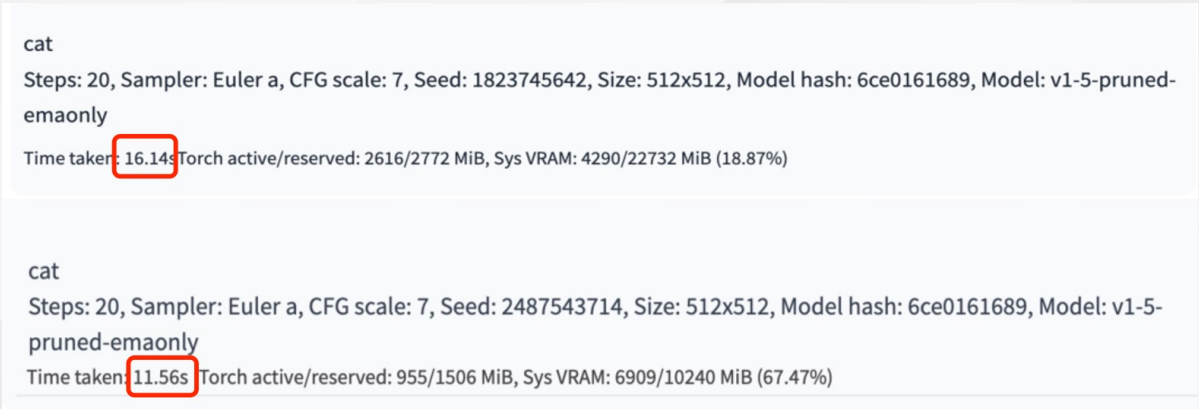


Stable Diffusion 是一个多模型组成的扩散 Pipeline，主要由三个部分组成：变分自编码器 VAE、U-Net 和文本编码器 CLIP。推理耗时主要集中在 UNet 部分，我们选择对这部分进行模型优化，加速推理速度

TACO: 优化 Stable Diffusion 推理性能

1. 下载 A10 GPU 优化的 stable-diffusion-v1.5 UNet 模型文件到CFS
2. 下载 sd_v1.5_demo 镜像并上传到TCR，该镜像里的 Web UI 修改了模型加载代码，UNet 部分会加载独立优化模型。
3. 将 sd_v1.5_demo 镜像服务部署在 TKE 上：按前述步骤进行操作，其中替换镜像为 sd_v1.5_demo，并额外为 UNet 优化模型创建 CFS /data 目录和PV&PVC
4. 同样参数配置下，生成10张猫的图片，优化前推理耗时16.14s。加载 TACO 优化的 UNet 模型后，10张图片仅耗时11.56s，端到端性能提高30%





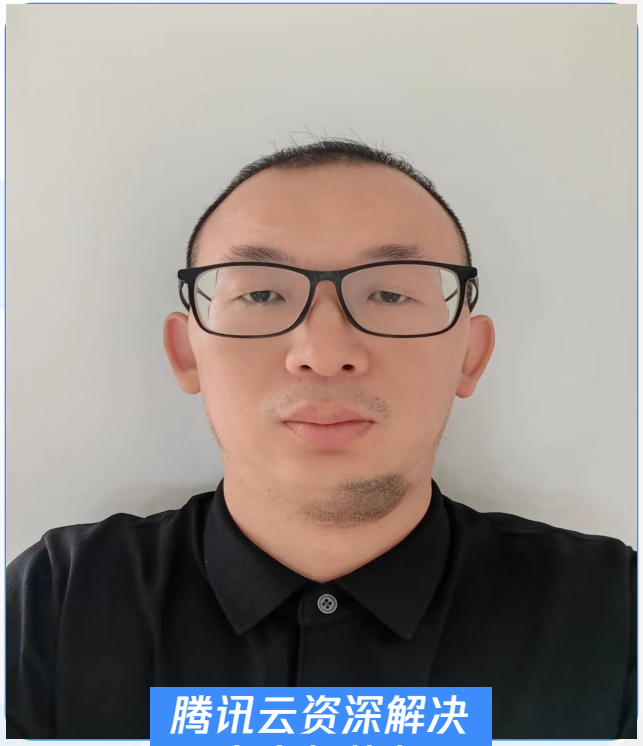
腾讯云

“中小企业”在线学堂

基于TI-ONE平台的SD推理加速实践

目录 Menu

讲师介绍



腾讯云资深解决方案架构师

孔波

十余年云计算从业经验，擅长云原生架构，混合云架构等，主要为腾讯云SaaS客户提供解决方案设计与落地咨询，目前关注AIGC等新技术推广落地

SD应用场景及挑战

应用场景

挑战

SD 基础模型

- **图像编辑**

去水印，提高分辨率，特定滤镜

- **文字生成图像**

根据文字prompt生成创意图像

Controlnet 控制模型

- **功能图像生成**

根据指定要求生成营销类海报，模特图，LOGO，建筑风格设计等

- **文字生成图像**

根据文字prompt生成特定要求和风格的图片

✓ **模型版本多**

开源模型多，同时需要根据业务做场景微调

✓ **生成速度慢**

SD基础模型往往生成一张高清图需要十几秒，加上Controlnet模型后，生成时长或达几十秒

✓ **业务流量不稳定**

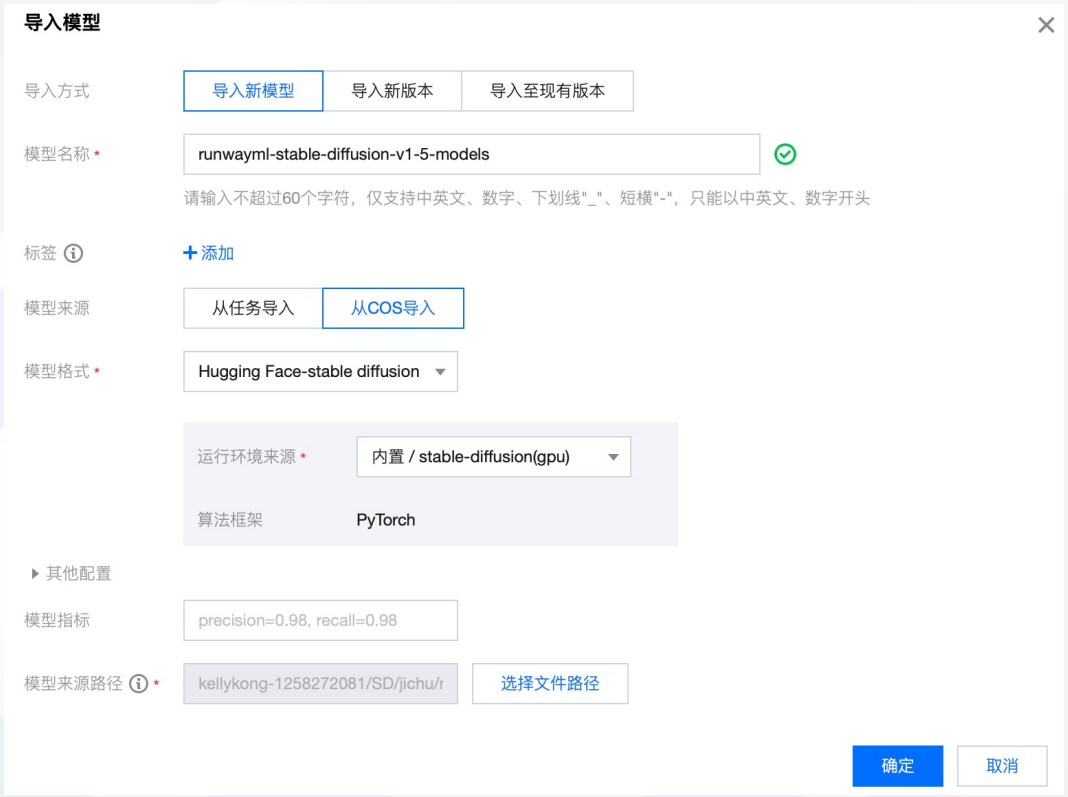
业务经常受营销活动推广影响，无法预测准确流量，GPU资源本身较贵

TI-ONE平台介绍


腾讯云TI平台
数据中心
数据集管理
个人标注
自动学习
场景化自动学习
作业配置
训练监控
模型评测
一键发布
训练工坊
Notebook
预置框架（含TI-ACC训练加速）
本地调试
TI-Kit任务提交
生命周期配置
Git存储库管理
任务式建模 Tikit
多模式训练任务创建
训练过程可视化
资源监控可视化
训练指标可视化
日志持久化
模型管理
模型仓库
模型导入
模型管理
模型优化
优化任务管理
推理优化（含TI-ACC推理加速）
模型服务
在线服务
模型/镜像服务化
手动/自动扩缩容
资源/调用数据监控
热更新
流量分配
在线测试
批量预测
任务管理
COS对接
对接云原生
cos对象存储
cbs云硬盘
cls日志服务
cm云监控
TCR 容器镜像服务
CAM访问管理
云API
云容器平台
腾讯云 IaaS 服务（计算+网络）

TI-ONE一键部署Stable Diffusion模型服务

1.TI-ONE-模型管理-模型仓库模块导入模型



2.TI-ONE-模型服务-在线服务-新建服务，选择刚导入的模型

3.耐心等待几分钟，TI-ONE-在线服务

“中小企业”在线学堂

TI-ONE一键部署Stable Diffusion模型服务

3.TI-ONE-在线服务-选择服务-调用

接口信息

接口调用地址：http://service-ldnm36ju-1308945662.gz.apigw.tencentcs.com:80/tione　/v1/models/m:predict

服务类型：HTTP

请求方法：POST

调用方式(命令行)：curl -X POST http://service-ldnm36ju-1308945662.gz.apigw.tencentcs.com:80/tione/v1/models/m:predict -H 'Content-Type:application/json' -d ''

若服务开启了鉴权，请参考文档 指引调用

调用方式(在线测试)

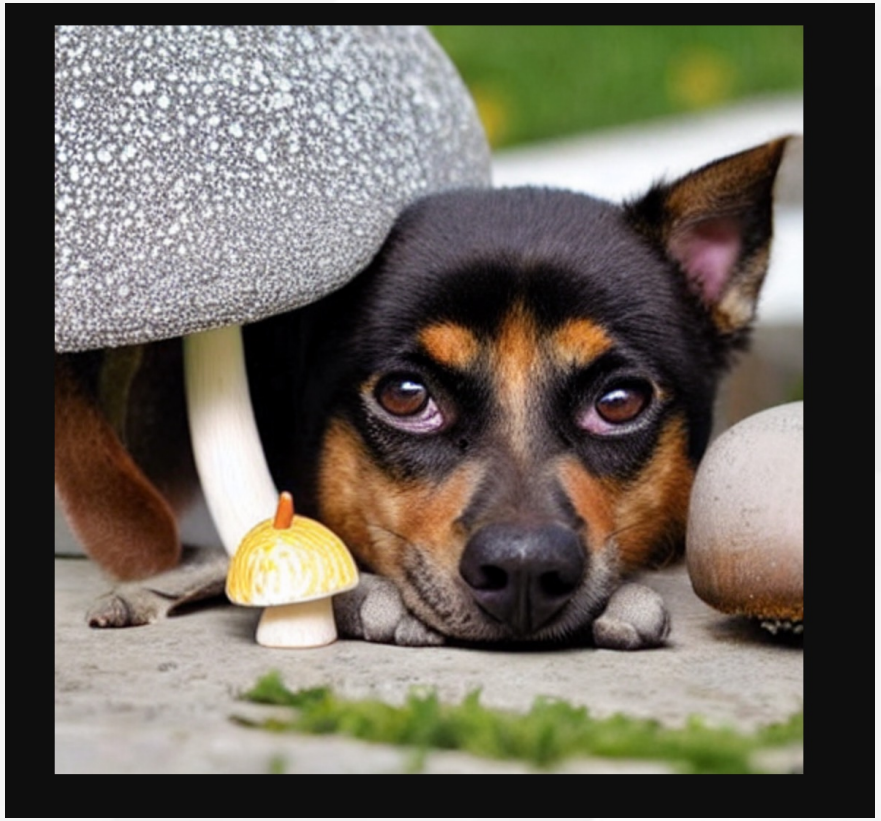
请求体(Request Body 600KB 内)

```
1  {"prompt": "a cute dog sleeing with a mushroom"}
```

请求响应(Response)

```
1  Status: 200 OK
2  Connection: keep-alive
3  Content-Length: 119
4  Date: Fri, 19 May 2023 11:30:08 GMT
5  Server: openresty/1.21.4.1
6  X-Api-Id: api-2g1w7sw2
7  X-Api-Requestid: 3b88d715da85303bd9cbebb668f416db
8
9  {
10     "image": "https://demo-1256580188.cos.ap-guangzhou.
   myqcloud.com/sd-demo/8832e4f6-e670-4e9a-bf98-8da1e7680963.png"
11 }
```

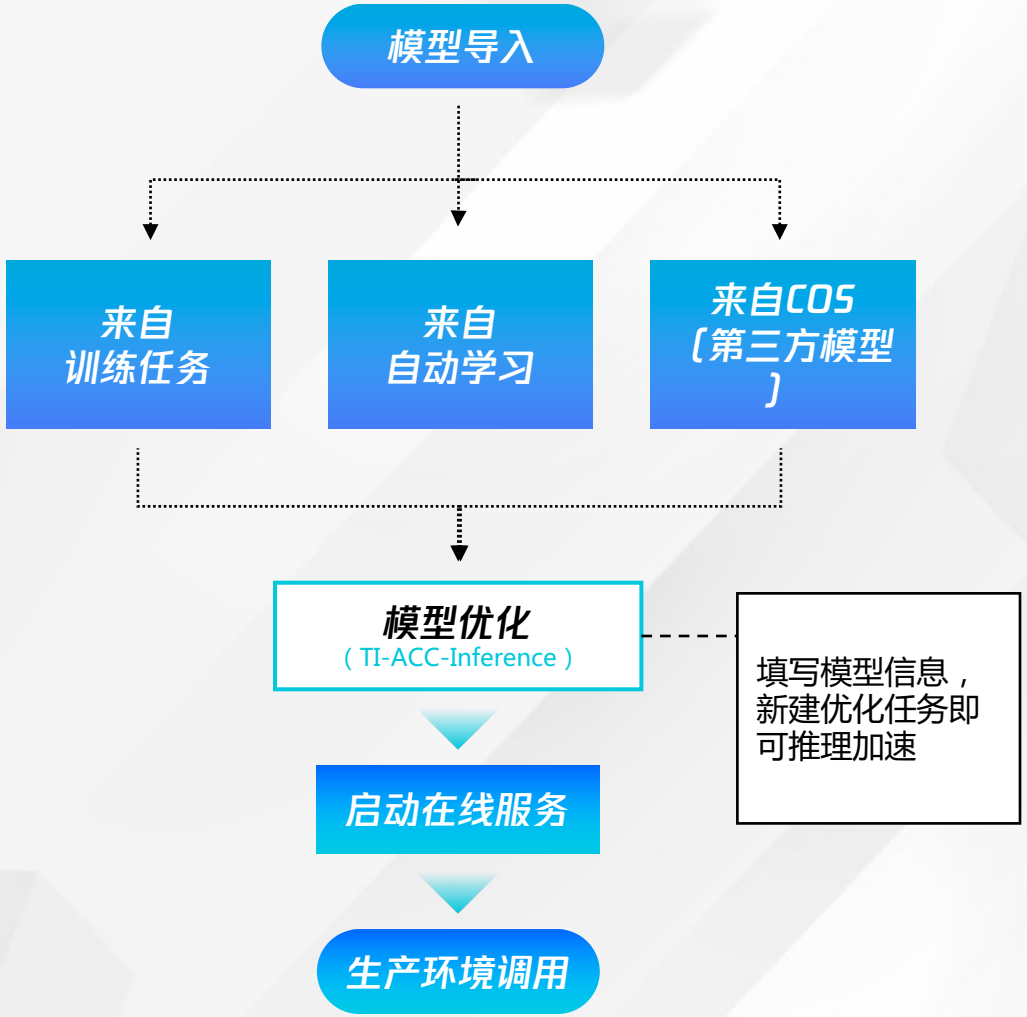
发送请求



TI-ACC-Inference加速推理介绍

功能概述

TI-ACC推理加速在CV、NLP、推荐等模型推理场景中，用户仅需新建一个优化任务即可进行推理加速优化，可帮助客户模型推理时，显著节省推理时间和计算成本。



适用客户

广泛集中在泛互VIP、KA以及在线教育等领域

✓有算法研究员储备，业务需要高精度AI模型赋能

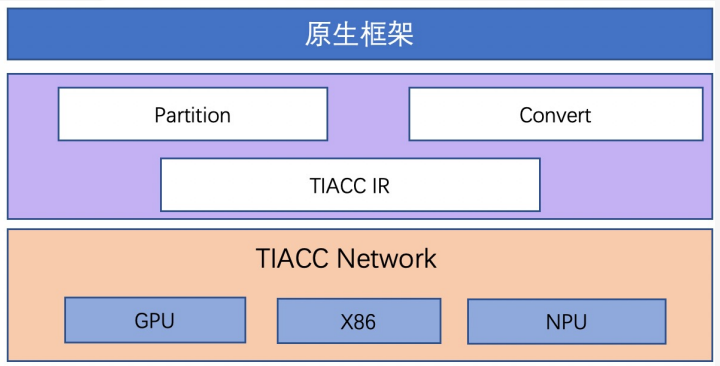
✓AI模型部署规模在10卡-数百卡

✓无专门研究AI加速的团队

TI-ACC-Inference加速推理介绍

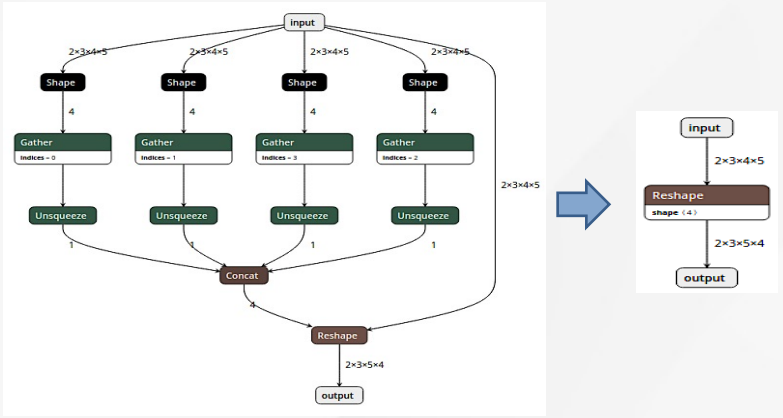
子图切分

基于原生框架，通过子图切分替换将加速能力嵌入原生框架，保证了接口与原生接口一致，同时保证了加速能力的通用性。



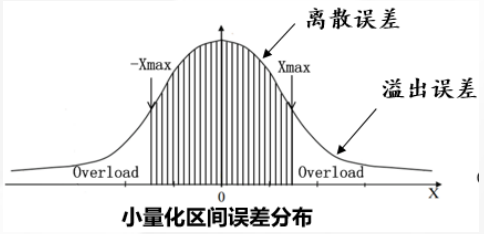
常量折叠

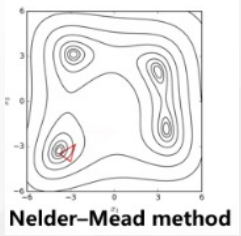
离线常量折叠，对常量部分进行预计算，降低运行时计算量。



Post Train

基于小量样本数据，统计数值分布，迭代求解各层最优量化区间



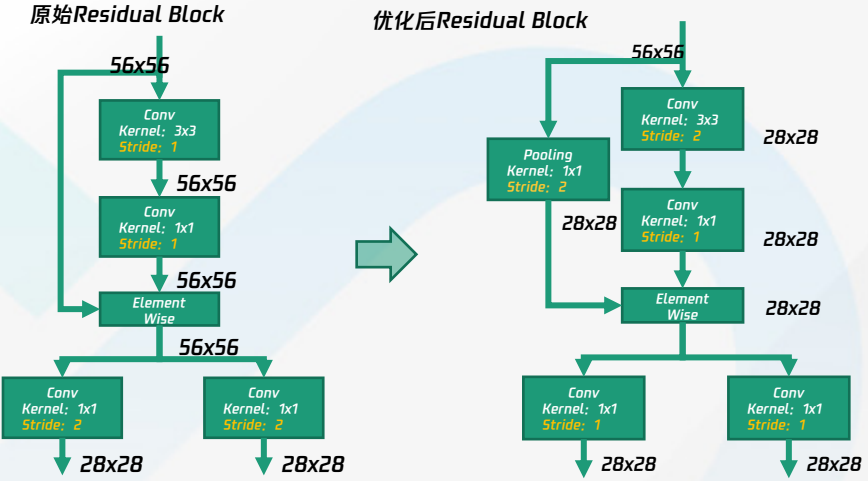


$$\sigma^2_{q(total)} = \sum_{q=1}^{M} \int_{X_{\min}+\Delta\cdot(q-1)}^{X_{\min}+\Delta\cdot q} [x-(X_{\min}+\Delta\cdot q-0.5\Delta)]^2\, p(x)dx \quad \text{离散误差}$$

$$+\int_{X_{\max}}^{\infty} [x-(X_{\max}-0.5\Delta)]^2\, p(x)dx$$

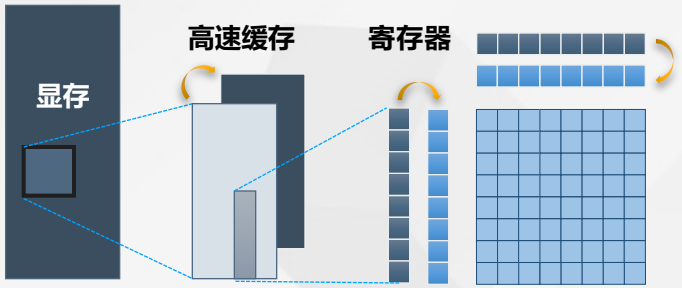
$$+\int_{-\infty}^{X_{\min}} [x-(X_{\min}+0.5\Delta)]^2\, p(x)dx. \quad \text{溢出误差}$$

图优化



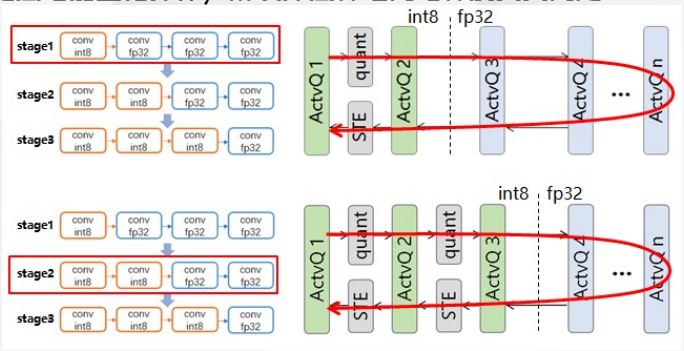
针对架构深层性能优化

- ✓ 细致Kernel设计，提高GPU Occupancy
- ✓ Precompute Offset通过GPU Const Memory提高访存效率
- ✓ Double Buffering掩盖延迟
- ✓ ILP，外积累加的方式实现指令级并行



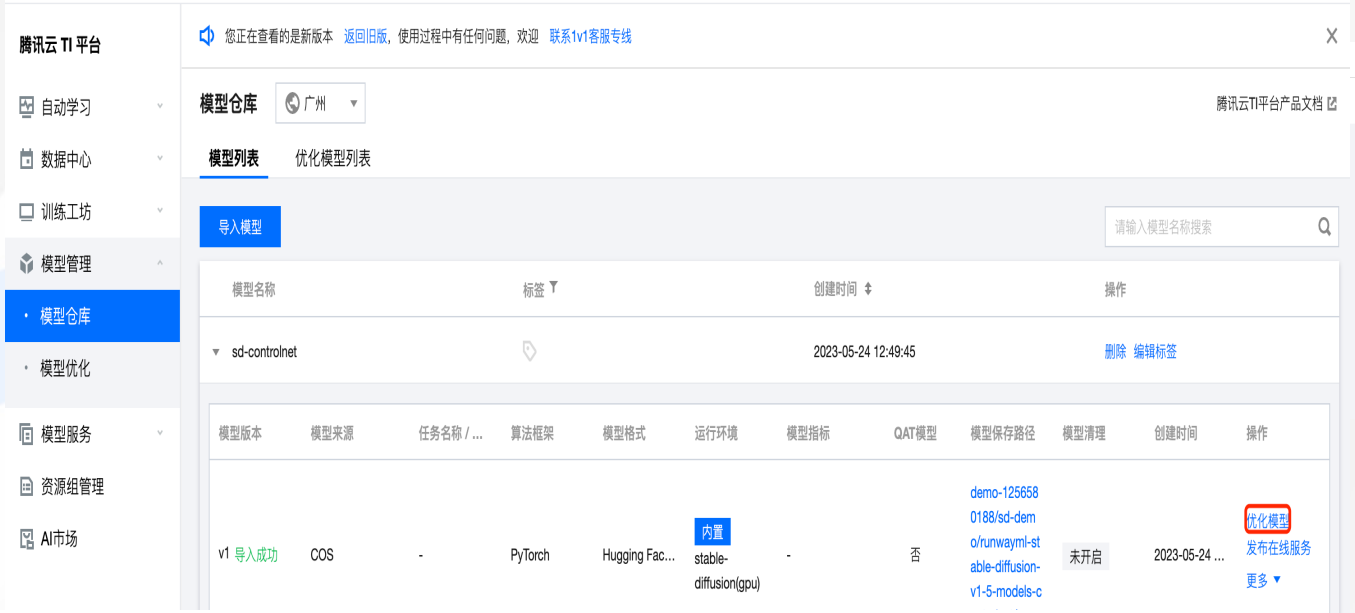
QAT 分阶段量化策略

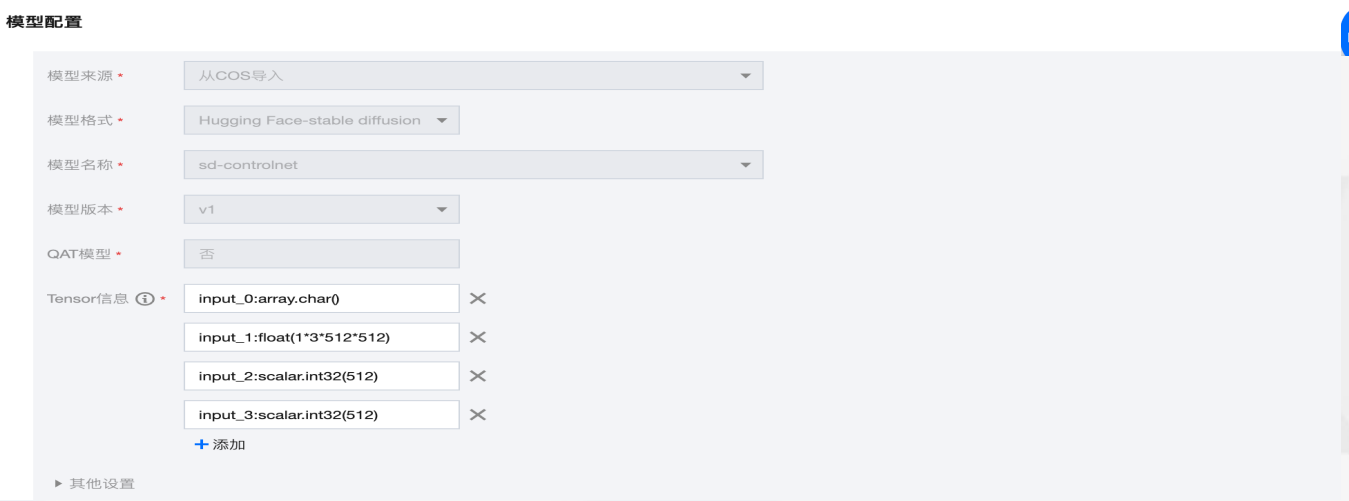
逐层增加量化操作，解决深层神经网络梯度失真问题



SD+Controlnet推理加速实战

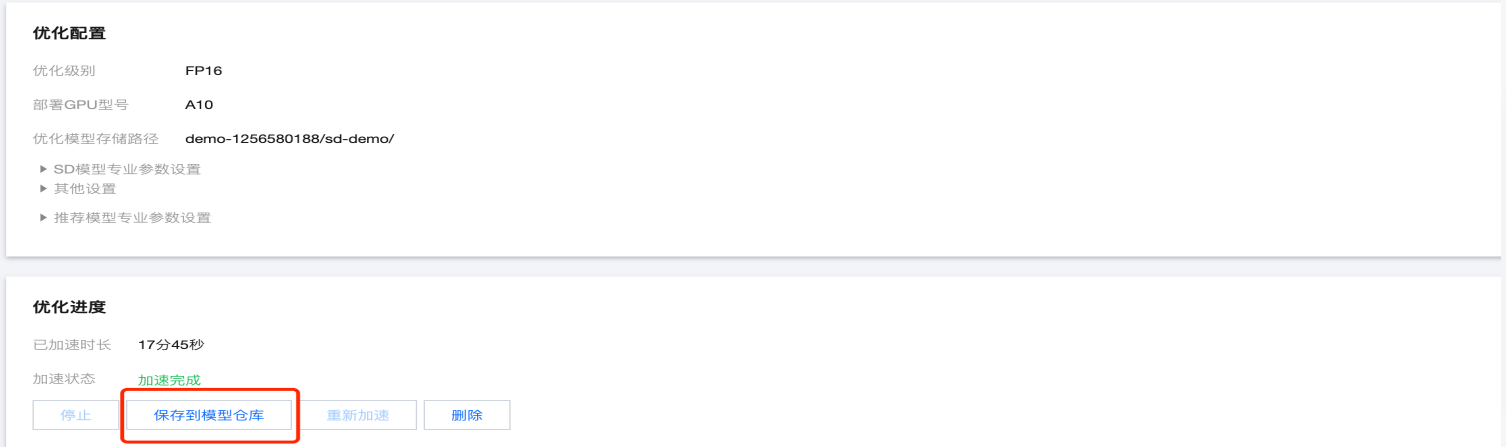
1.TI-ONE-模型管理-模型仓库-优化模型



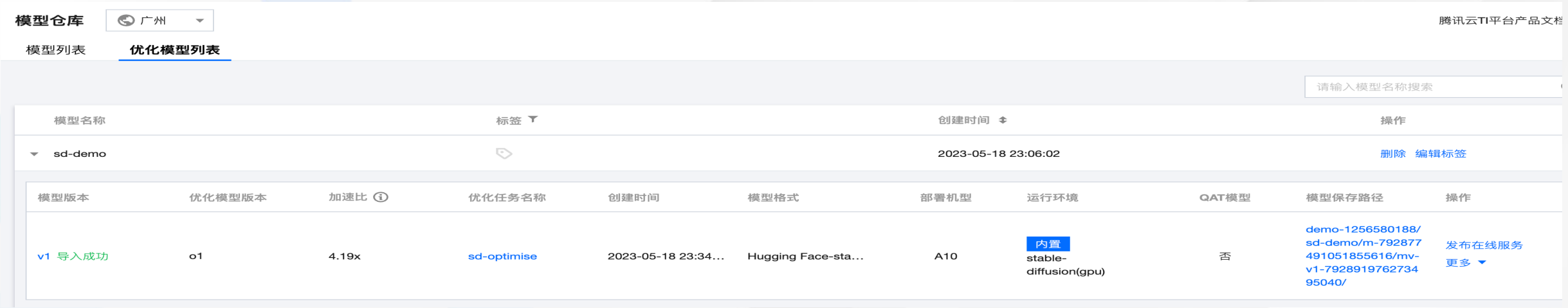


SD+Controlnet推理加速实战

2.TI-ONE-模型管理-模型优化-保存模型仓库

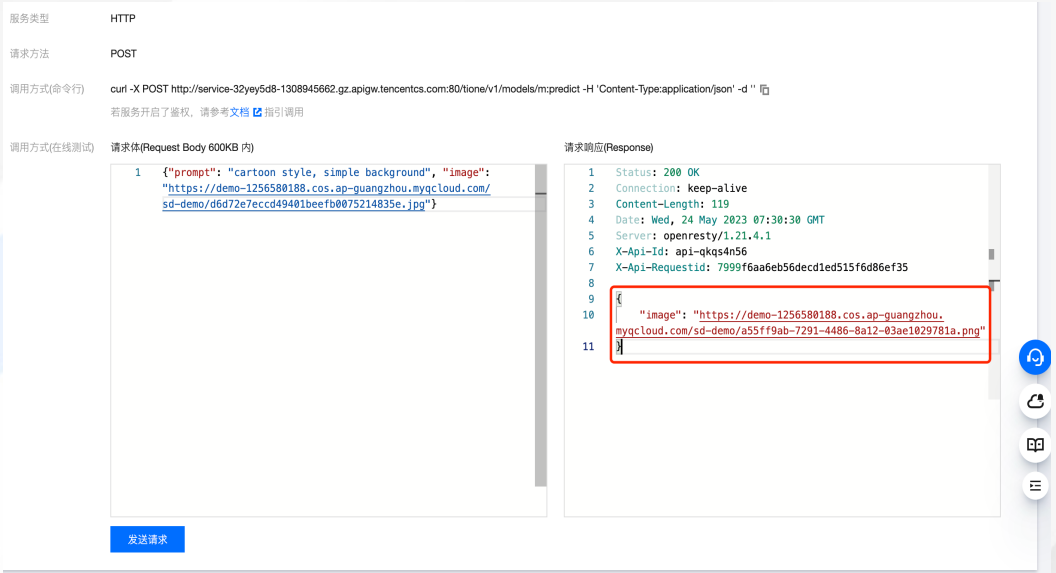


3.TI-ONE-模型管理-模型仓库-优化模型列表-发布在线服务



SD+Controlnet推理加速实战

4.TI-ONE-在线服务-选择服务-调用



原始图像

生成图像

ControlNet评测

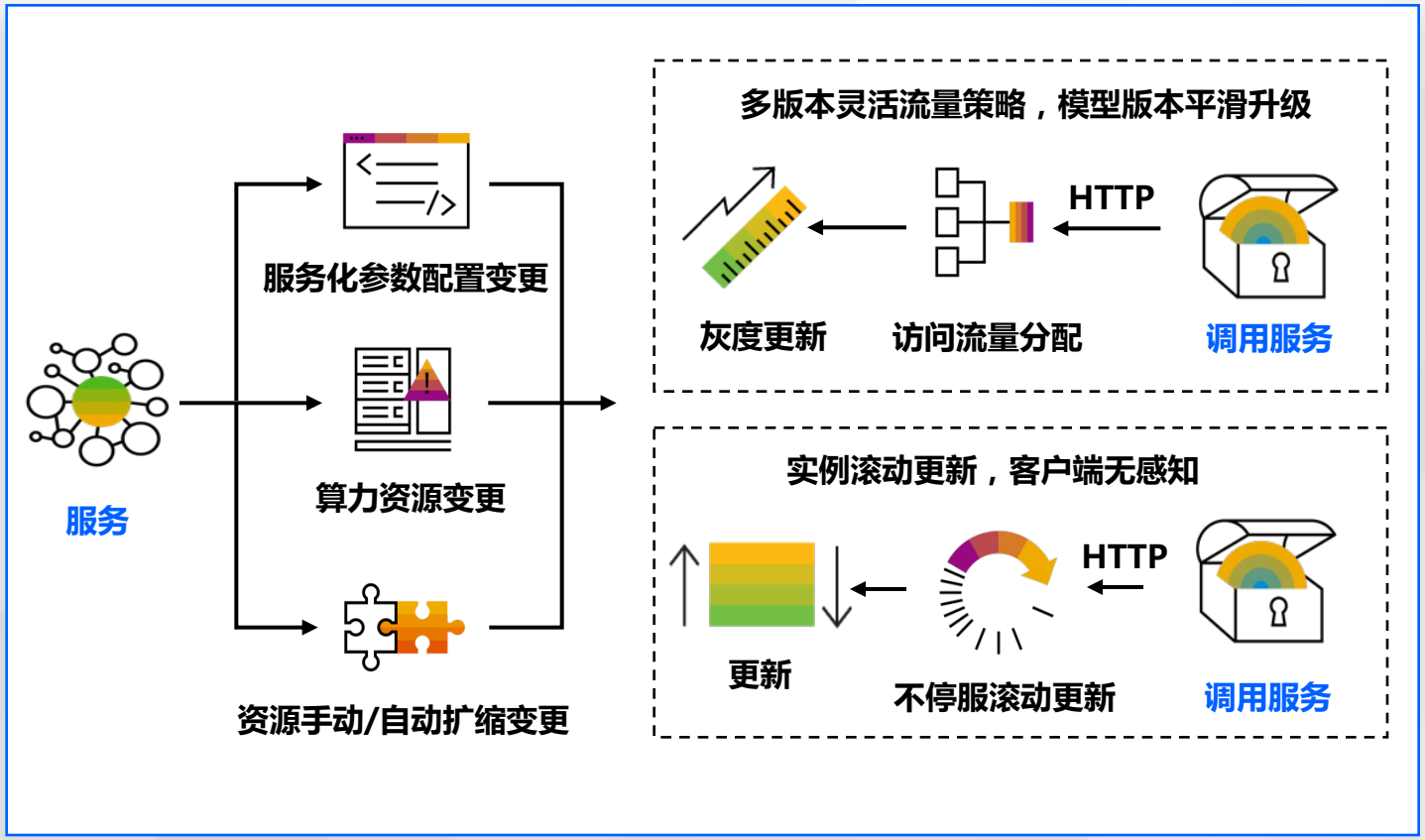
runwayml/stable-diffusion-v1-5 + Canny ControlNet

图片输出尺寸: 512x512

ControlNet 评测(A10) (torch1.12.1)	huggingface fp32	huggingface fp16	TI-ACC
耗时	13811ms	6463ms	3207ms

利用TI-ONE部署弹性伸缩SD推理服务

- ✓ 支持服务不停服滚动更新，确保服务持续在线可用
- ✓ 支持服务多版本管理与流量分配，支撑灰度测试等服务运营管理场景
- ✓ 支持实例级别的手动/自动算力资源扩缩容，高效利用集群算力
- ✓ 支持服务算力资源使用情况、服务调用情况的数据监控能力
- ✓ 支持服务在线调用测试



利用TI-ONE部署弹性伸缩SD推理服务

TI-ONE-在线服务-选择服务-扩缩容

感谢观看！
Thank you