材料科研软件平台使用文档

平台简介

平台提供了基于国产硬件的大模型训练优化框架、材料领域预训练数据及任务指令知识构建、训练优化及工具集成、协同开发服务环境、材料大模型推理及部署的全生命周期管理服务，为材料大模型训练与持续演化提供支撑。平台提供丰富的材料大模型训练组件、镜像、数据集、模型等，支持主流的深度学习框架及多种模型的开发方式。同时支持资源灵活调度、团队协作开发、版本管理等功能。

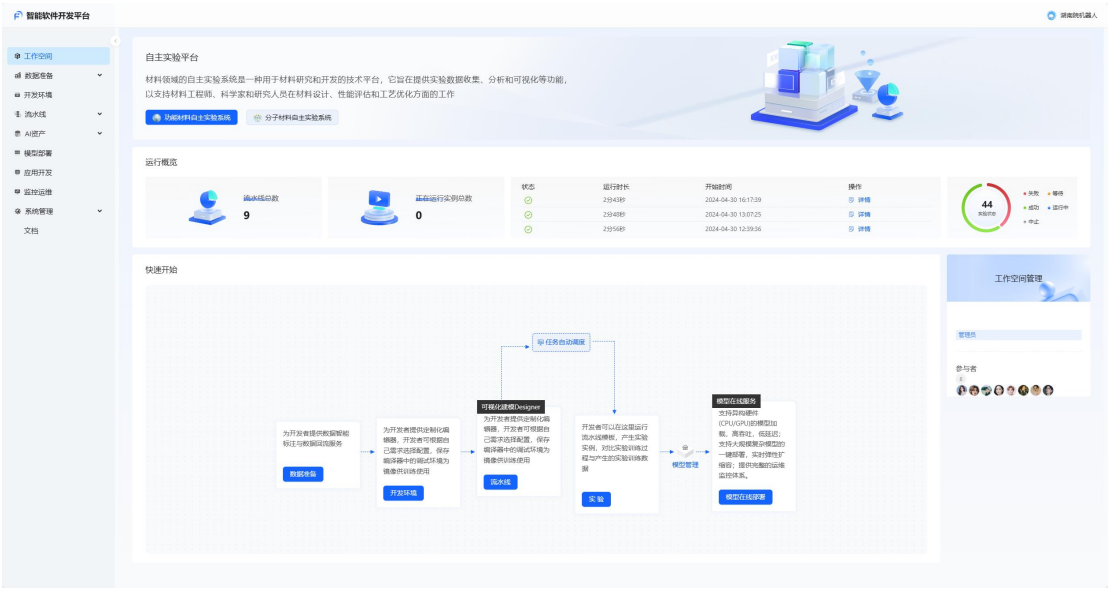


图 1 智能软件研发平台

模块介绍

1. 数据标注

数据标注平台是一个为机器学习和人工智能模型训练提供数据标注服务的工具，它能够帮助用户高效、准确地标注数据，并提供了一系列功能来管理、控制标注质量、协作和定制化，从而提高模型训练的效率和准确性

1.1 创建标注项目

点击“数据准备”，再点击“数据标注”，打开数据标注页面，如图所示：

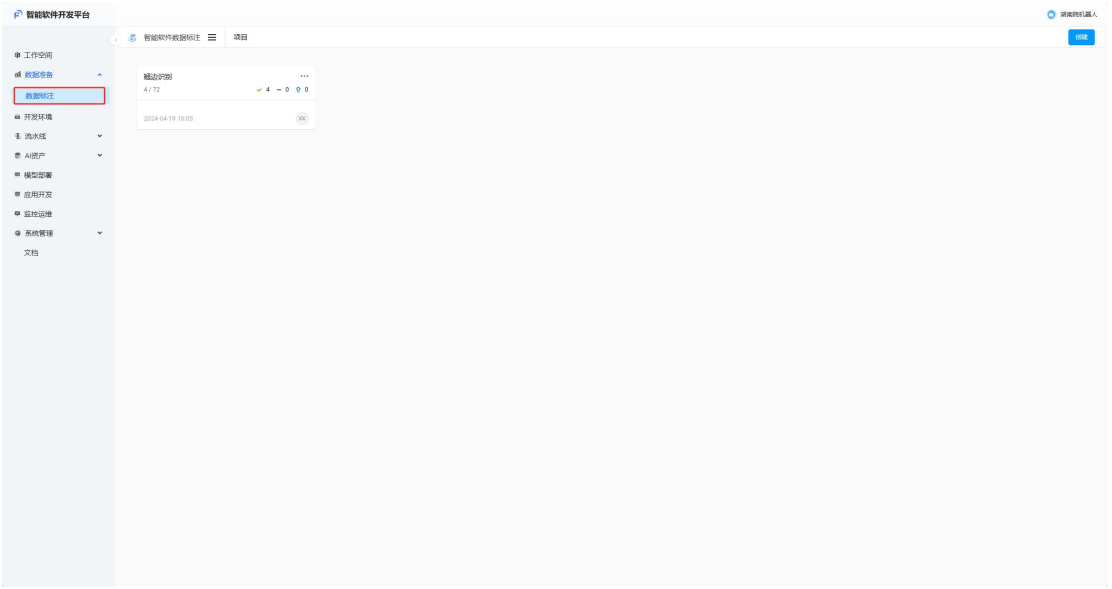


图 2 数据标注

点击“创建”创建项目，填写项目名称和项目描述，然后点击“数据导入”上传标注文件，最后点击“标注配置”进行数据标注配置

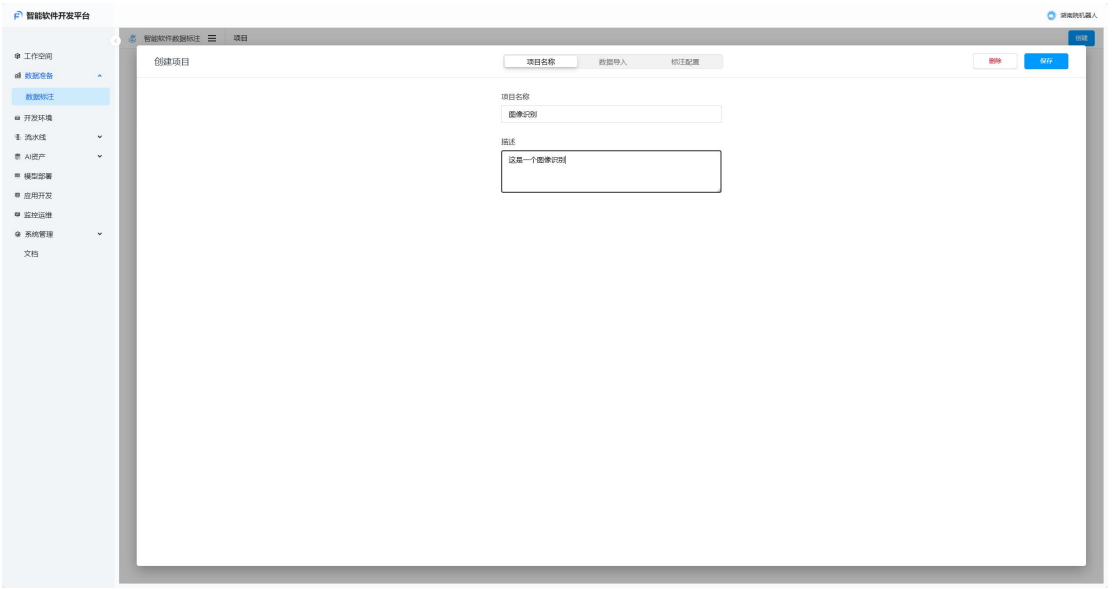


图 3 数据标注创建项目

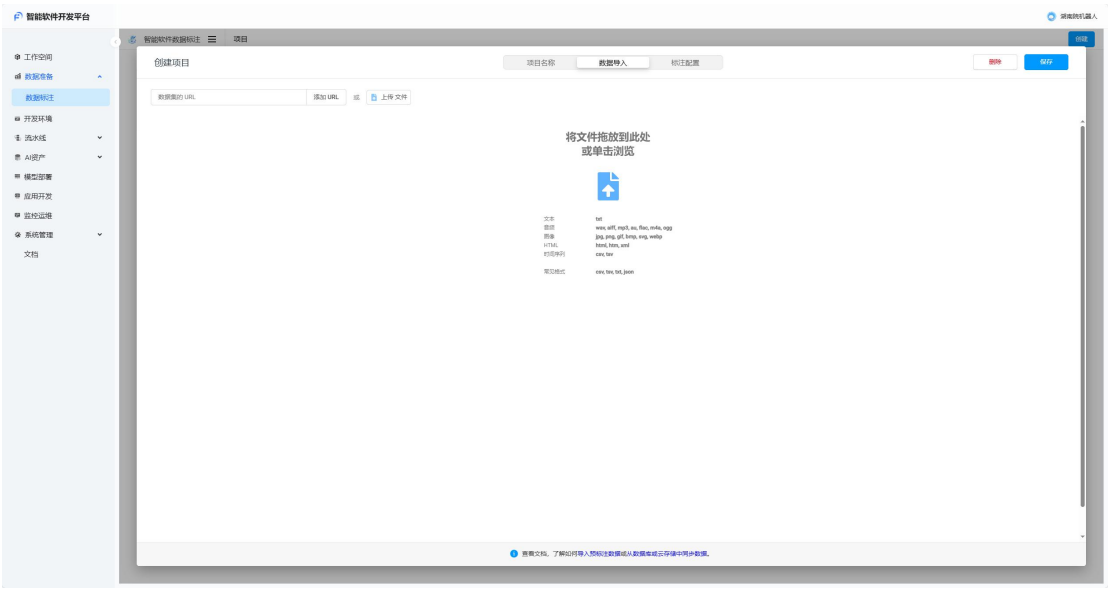


图 4 数据标注上传标注文件

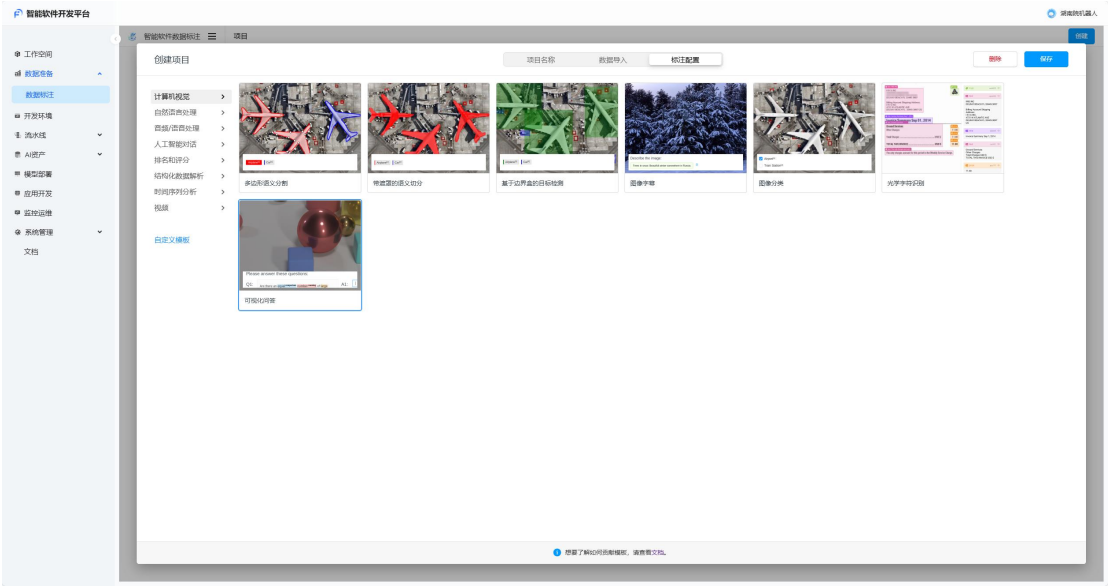


图 5 数据标注选择标注模板

1.2 标注数据

标注数据和标注配置好以后，点击“所有的标注任务”对数据进行标注

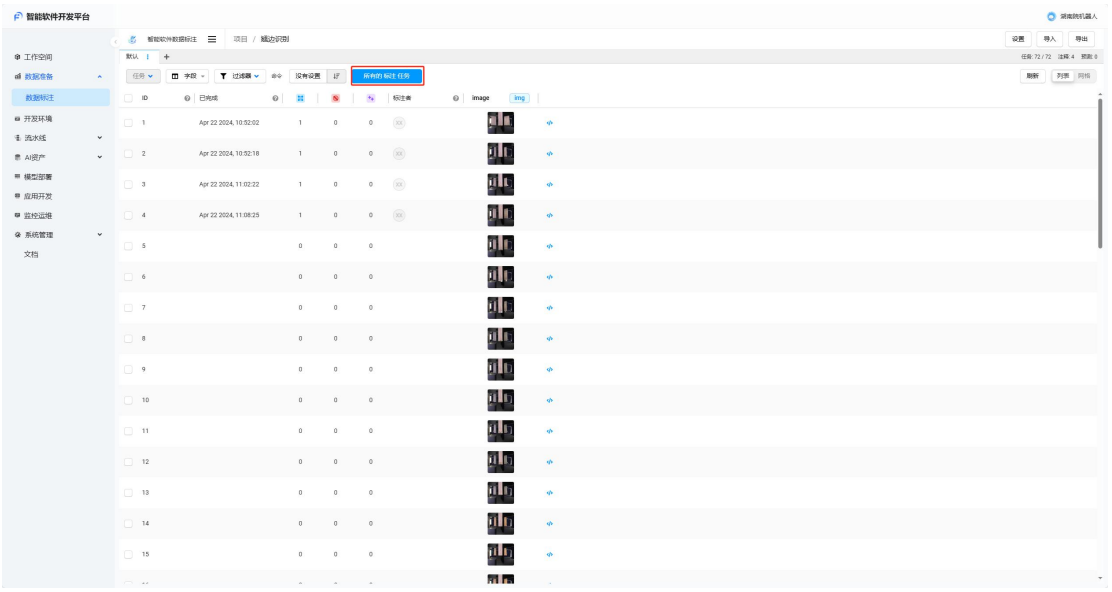


图 6 所有的标注任务

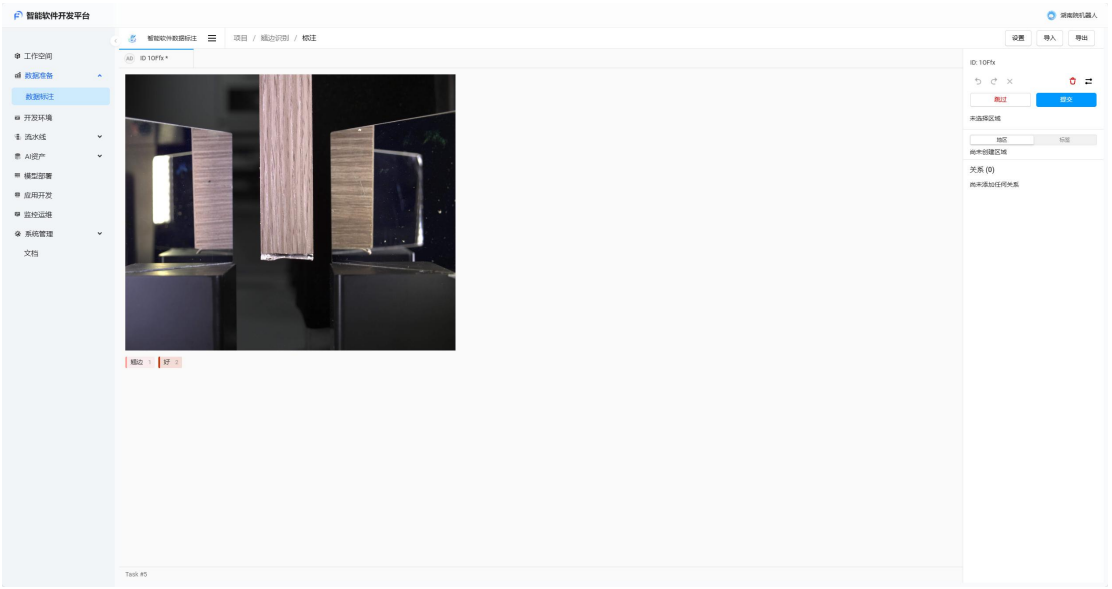


图 7 标注

1.3 标注结果导出

数据标注完成后，点击“导出”对标注的数据进行导出，标注的后数据集可以上传到平台的数据集共模型训练使用。

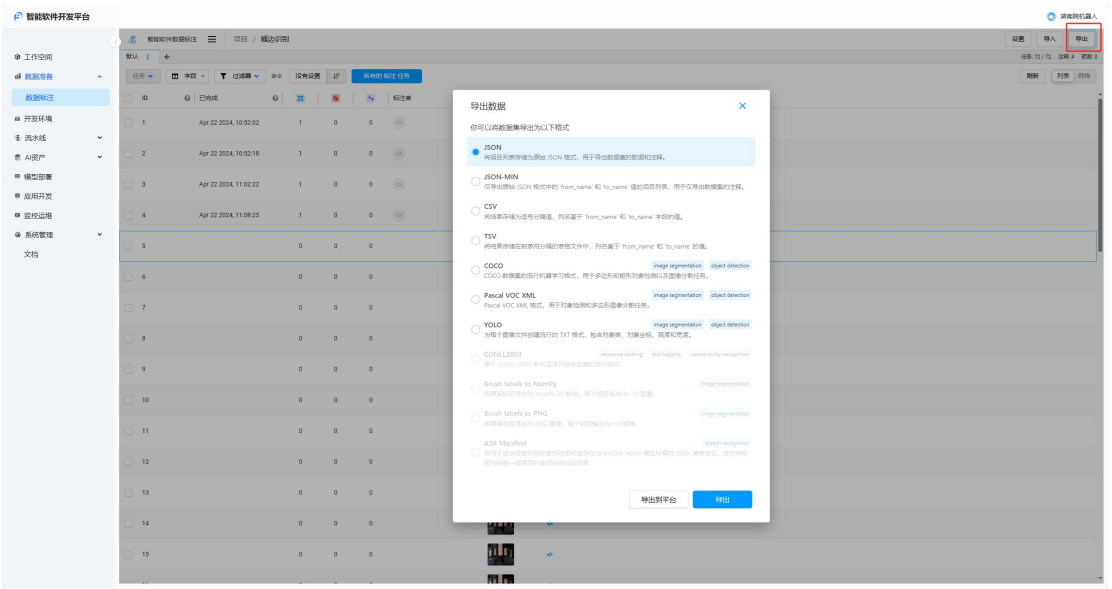


图 8 导出已标注数据集

选择导出数据格式后，点击“导出到平台”，选择需要导出至的平台数据集后，填写版本，描述，标注的后数据集可以上传到平台的数据集。

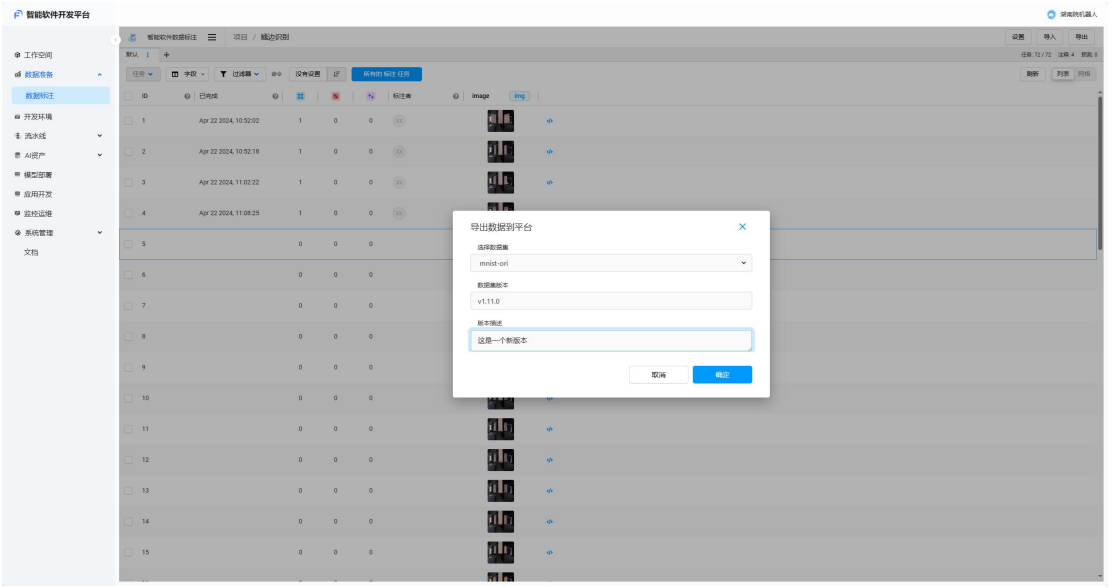


图 9 导出至平台数据集

选择导出数据格式后，点击“导出”，已标注数据集导出至本地。

2. 开发环境

开发环境为用户提供了一个在线的代码编写、代码调试环境，用户可以在开发环境编方便的编写代码和调试代码。

2.1 开发环境使用

点击“开发环境”，启动开发环境编辑器，如图所示：

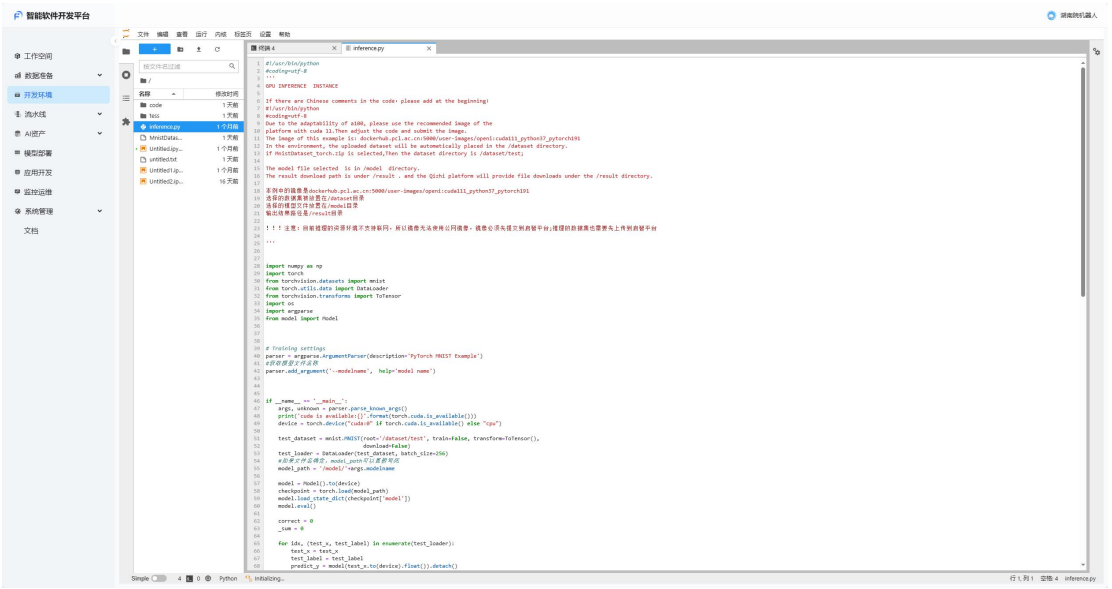
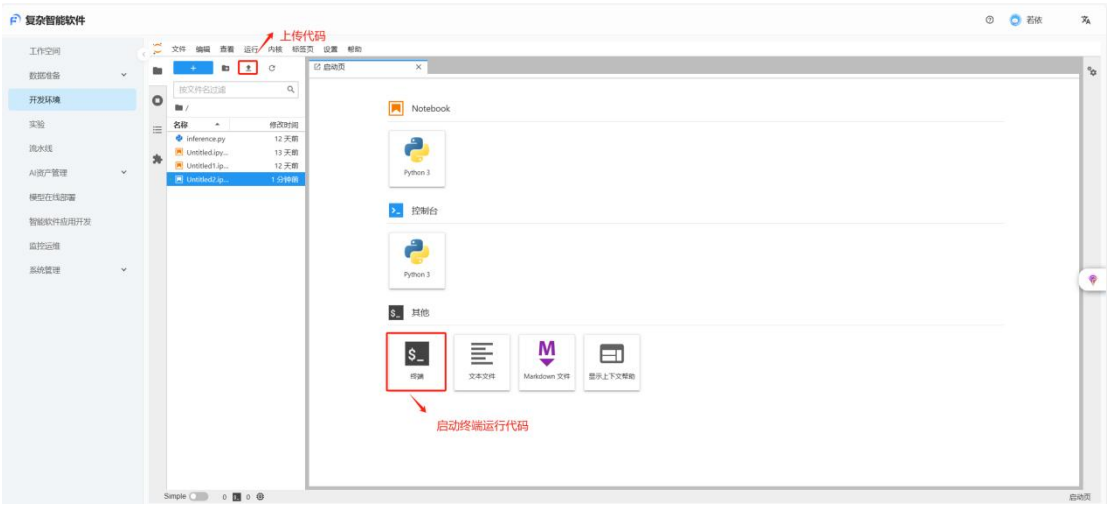
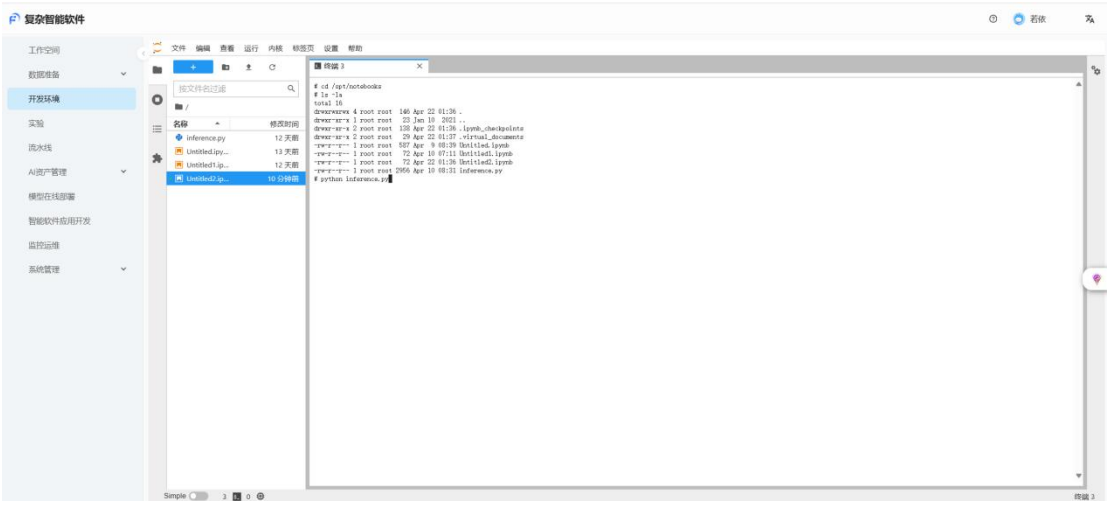


图 10 开发环境

用户可以上传代码文件或者数据集模型文件，上传的文件保存在终端的/opt/notebooks 目录，点击终端进入/opt/notebooks 目录可以运行代码。





3. 流水线

流水线为用户提供了数据处理、模型训练、模型测试、数据导出等丰富的组件，用户可以根据业务场景的不同，通过可视化的方式快速构建 MLOps 流水线。

3.1 新建流水线

点击“流水线模板”菜单，点击“新建流水线”可以创建流水线。

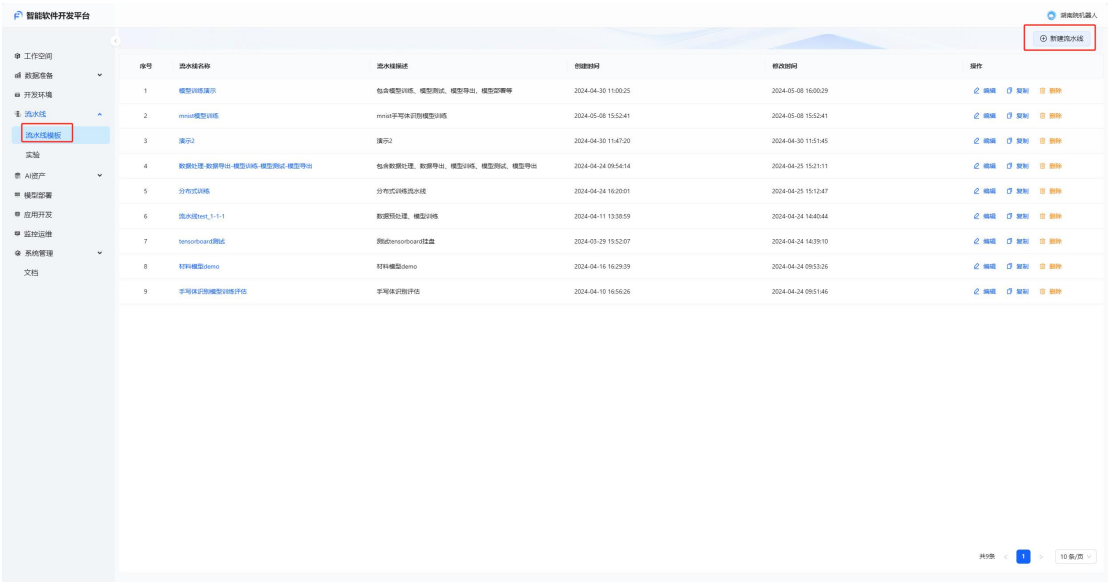


图 11 新建流水线

3.2 复制流水线

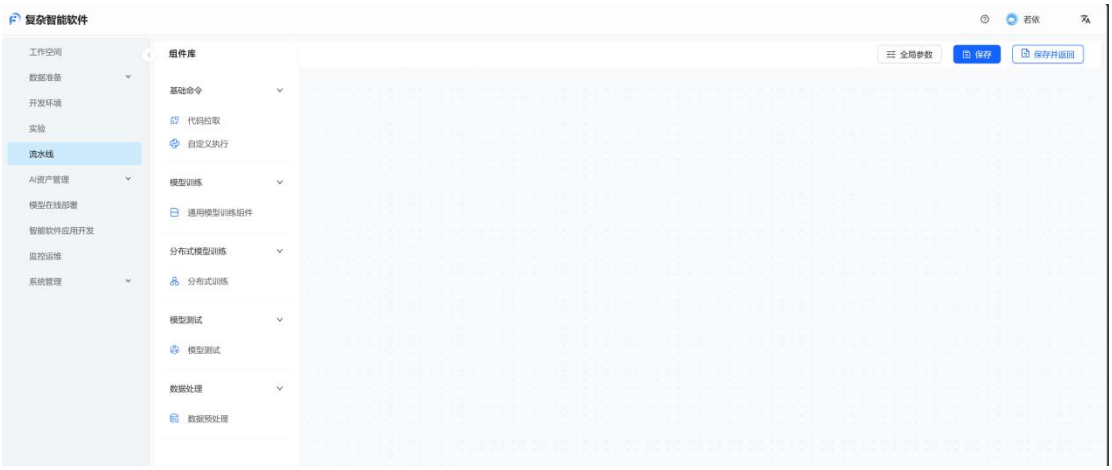
用户也可以通过现有的流水线，点击“复制”快速复制出一条流水线，复制出的流水线名称为“原流水线名称_copy”

3.3 编辑流水线基本信息

点击“编辑”按钮，可以对流水线进信息进行编辑

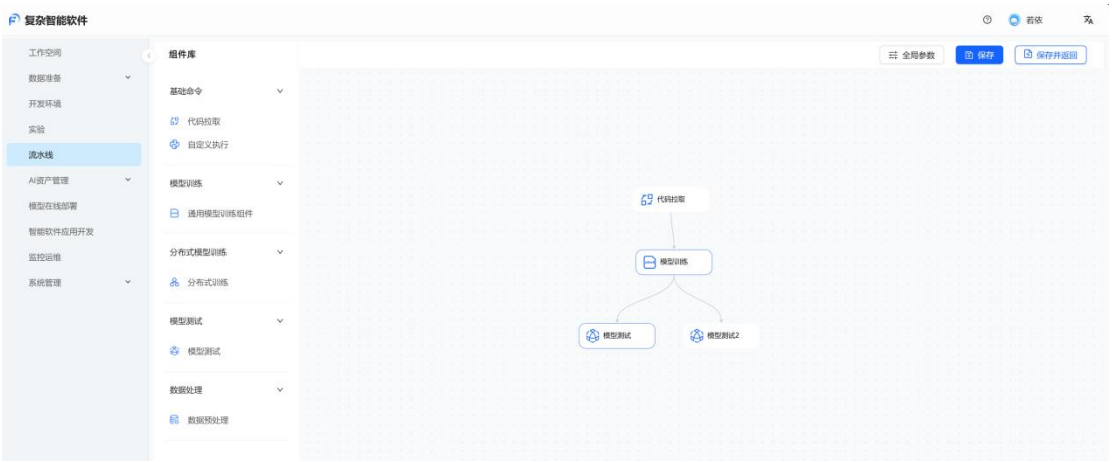
3.4 配置流水线

在流水线列表中点击流水线名称，可以进入配置流水线界面



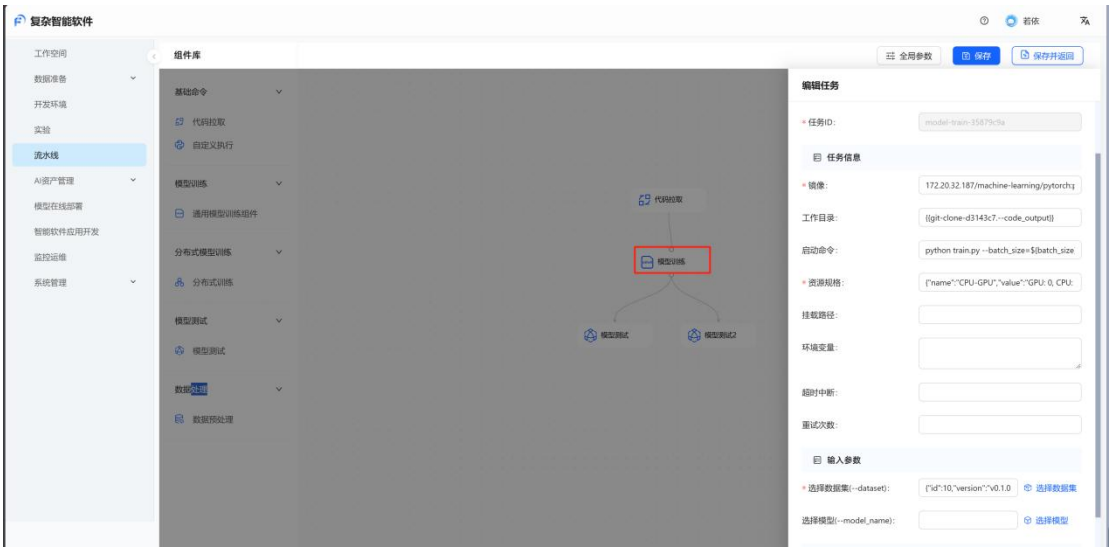
流水线拓扑配置

根据业务场景，从左侧流水线中拖组件到画布中，然后根据业务流程连接各个组件。



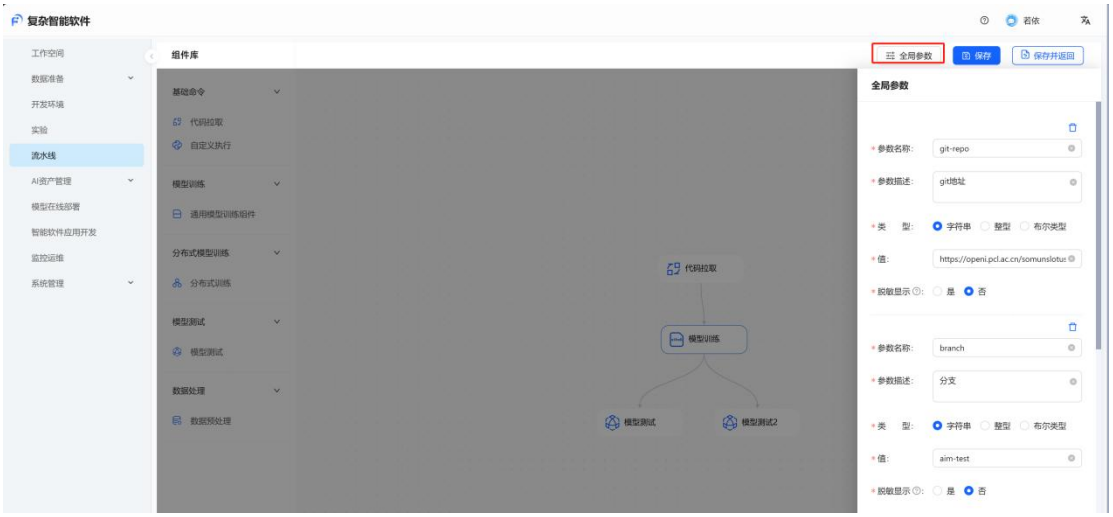
配置组件参数

将组件从组件库侧边栏拖到流水线画布编辑页面后，双击组件，既可以弹出组件需要配置的参数



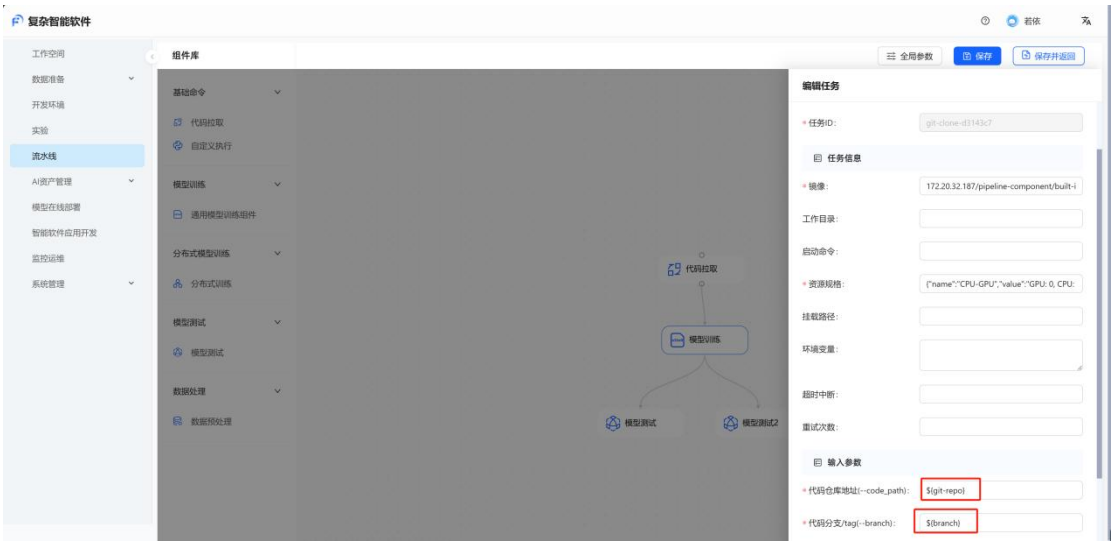
添加流水线全局参数

点击“全局参数”，在弹出的框中填写全局参数的需要填写的信息，信息填写完毕点击“保存”，全局参数即可保存。



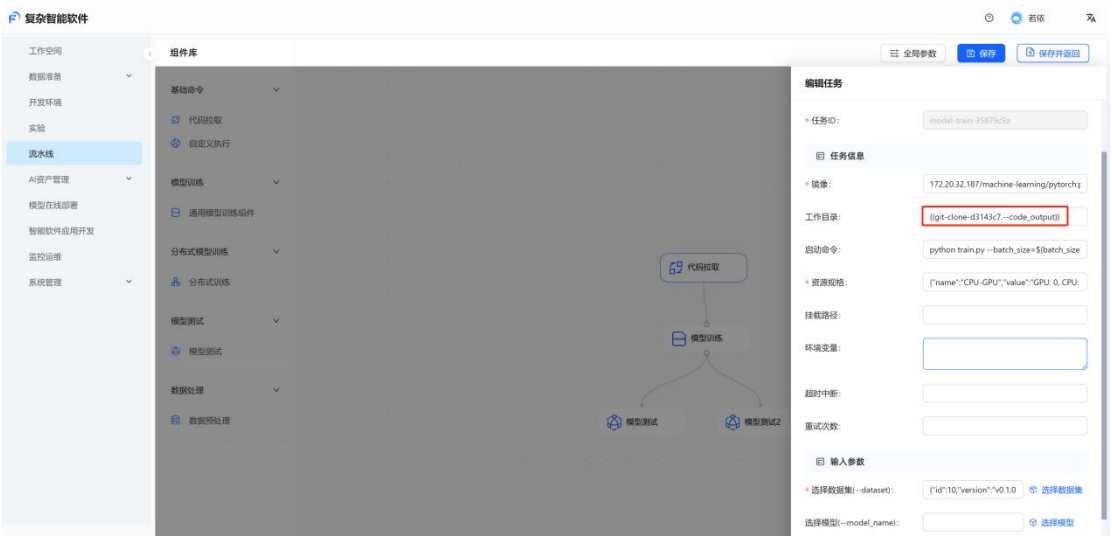
引用流水线全局参数

流水线全局参数设置好以后，可以使用${参数名称}或者$参数名称进行引用，如下图所示。



引用组件产物输出

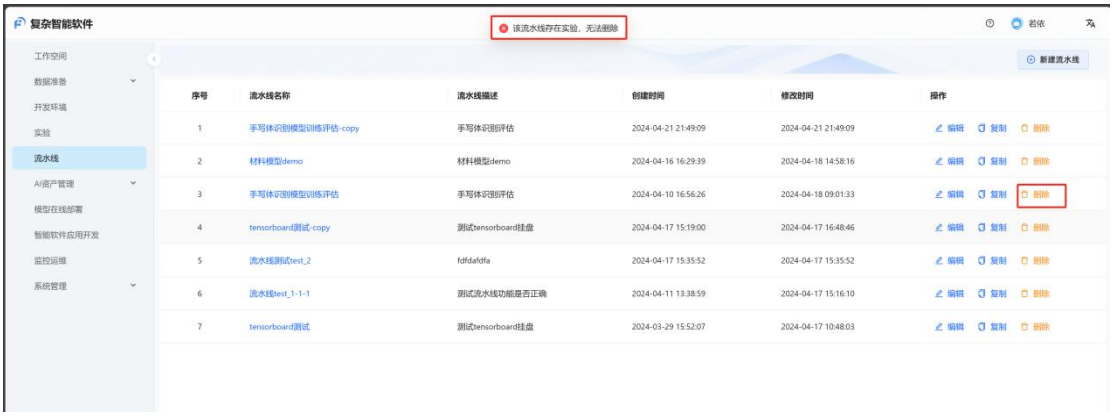
流水线中通常需要引用其他任务的输出结果，引用格式：{{task_id.param_name}}，如下图所示：



其中“git-clone-d3143c7”是代码拉取的 task_id, --code_output 是代码拉取组件的输出参数名称。

3.5 删除流水线

流水线如果有实验关联，该流水线不能被删除，需要先删除实验，再删除流水线，如果流水线没有实验关联，则可直接删除。



3.6 组件介绍

组件参数介绍

组件参数分为“基本信息”、“任务信息”、“输入参数”，“输出参数”四个部分。

基本信息：包含任务名称（在画布中显示的名称）、任务 id（组件从组件库侧边栏拖动到画布中自动生成，格式为“组件名称-随机字符串”）

任务信息：包含任务执行的通用信息。

镜像：任务运行需要的镜像，从平台提供的镜像列表中选择

工作目录：启动命令运行的工作目录

启动命令：如果启动命令为空，则使用镜像默认的启动命令，如果启动命令不为空，则会覆盖镜像的默认启动命令。启动命令后面可以带命令需要执行的参数

资源规格：任务运行需要的资源，目前支持 CPU/GPU 类型的规格，后续会支持其他的资源类型。

环境变量：任务运行时需要注入的环境变量，格式为 “key=value”，每行设置一个环境变量，多个环境变量需要换行设置。

超时中断：设置任务的最大运行时间，超过改时间任务没有运行完则任务自动失败，资源被回收。

重试次数：任务运行发生错误时自动重试的次数，最大重试次数为 5。

输入参数：输入参数分为两类，一类是挂载参数，一类是普通参数。所有的输入参数都会以参数的方式传给任务的启动命令，传入的参数名称为输入参数括号内的字符串。

挂载参数：包括选择数据集和选择模型两种参数，挂载输入参数会被挂载到任务执行的容器中。挂载的参数是以只读方式挂载的，防止用户误删除数据，用户使用的时候需要注意权限问题。

输出参数：所有的输出参数都会以参数的方式传给任务的启动命令。输出参数都会被挂载到任务执行的容器中，用户将结果写入挂载的路径中结果会持久化。

以模型训练为例，启动命令为“python train.py --batch_size=128 --epoch=2”，任务执行时真正的启动命令是”python train.py --batch_size=128 --epoch=2 --dataset=/worfklow-name/dataset --model=/workflow-name/model --model_output=/model”

说明：

输入参数：

--dataset=/worfklow-name/dataset 数据集挂载的路径，用户从--dataset 参数中可以读取悬选择的数据集。

--model=/workflow-name/model 模型挂载的路径，用户从--dataset 参数中可以读取悬选择的模型。

输出参数：

--model_output=/model 模型的输出路径，输出到该路径的模型会被持久化，任务结束可以查看和下载。

常用组件介绍

代码拉取：拉取组件运行所需要的代码，目前支持拉取 git 代码，不支持拉取其他方式的代码（svn 等），支持公有和私有 git 仓库的拉取，公有仓库使用 https 方式拉取代码，私有仓库使用 ssh 方式拉取代码。

通用模型训练：适用于规模不大的模型的训练，数据集和模型从平台的托管的数据集和模型中选取，只能使用单机的资源

分布式模型训练：适用于大模型的训练，数据集和模型从平台的托管的数据集和模型中选取，可以使用多级多卡资源，同时可以配置 worker 的数量。

模型测试：对模型进行测试，需要输入数据集和被测试的模型。

数据预处理：通用数据预处理，用户可以对选择的数据集进行处理，如果处理后的数据需要保存到数据集中，可以和数据集导出组件配合使用。

模型导出: 将模型训练输出的模型导出平台保存以便后续使用。如果没有注册过模型，需要先注册模型才能导出。

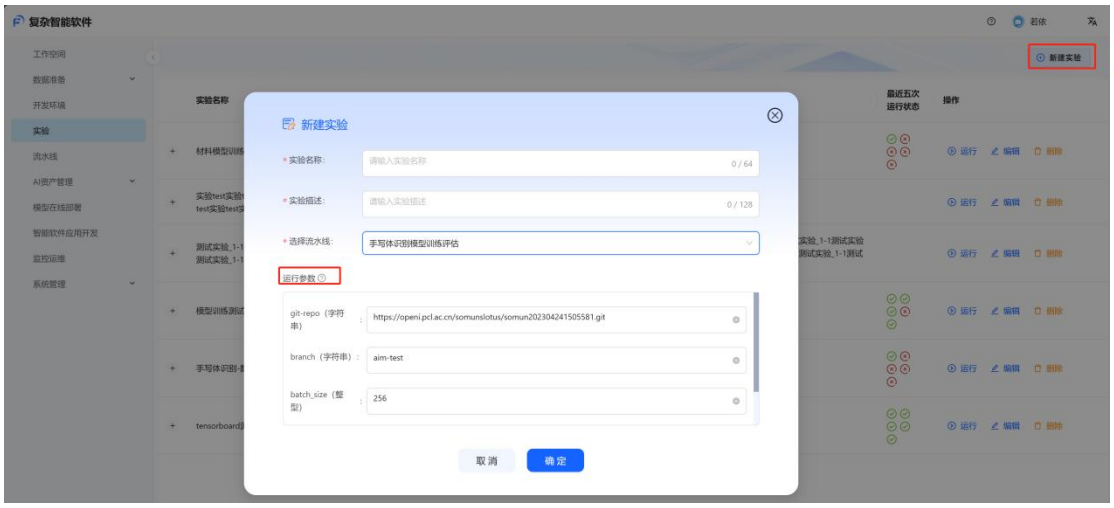
数据集导出：将数据处理后的结果导出到平台保存以便后续使用，如果数据集不存在，需要先创建数据集才能导出。

4. 实验

流水线配置完成后，用户可以选择流水线，动态调参进行实验。

4.1 新建实验

点击“新建实验”按钮，填写必须得参数，选择流水线，流水线选定后，流水线配置的全局参数会显示运行参数中，用户可以对初始值进行修改。



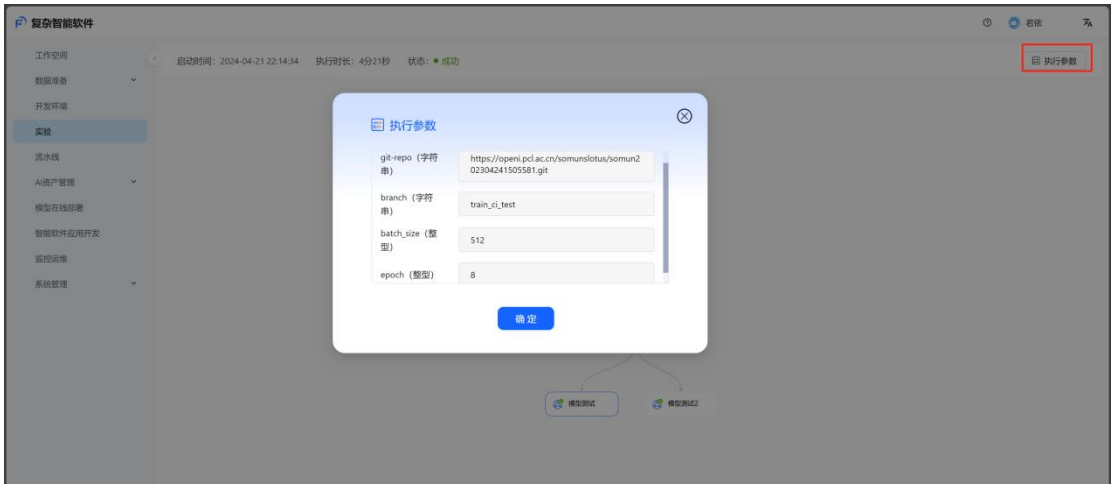
4.2 运行实验

新建实验成功后，点击“运行”，即可运行实验。

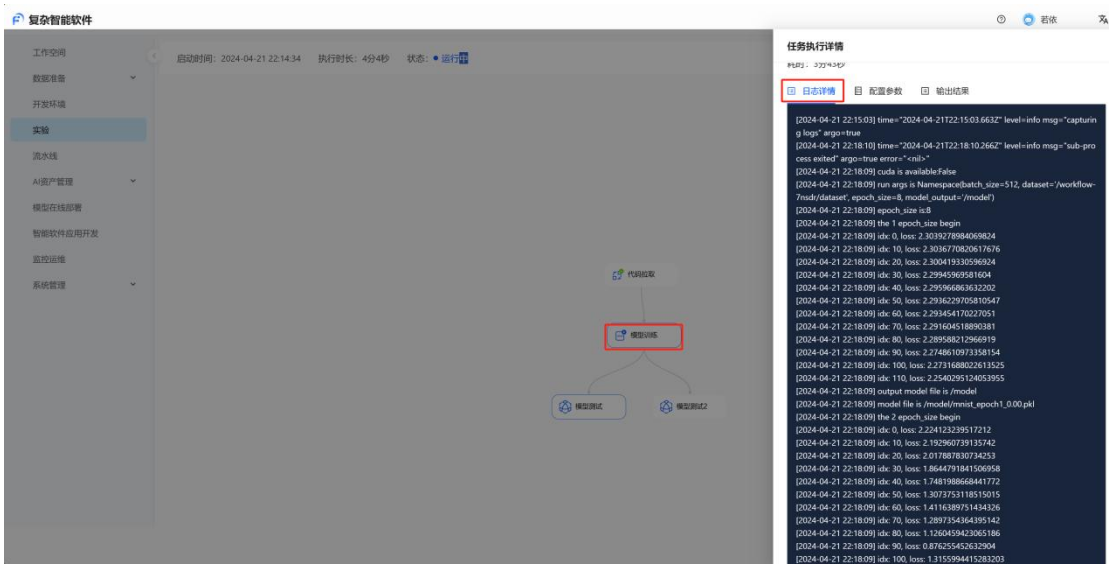
4.3 实验详情查看

实验运行后，点击实验运行的序号，可以进入查看实验运行的情况。

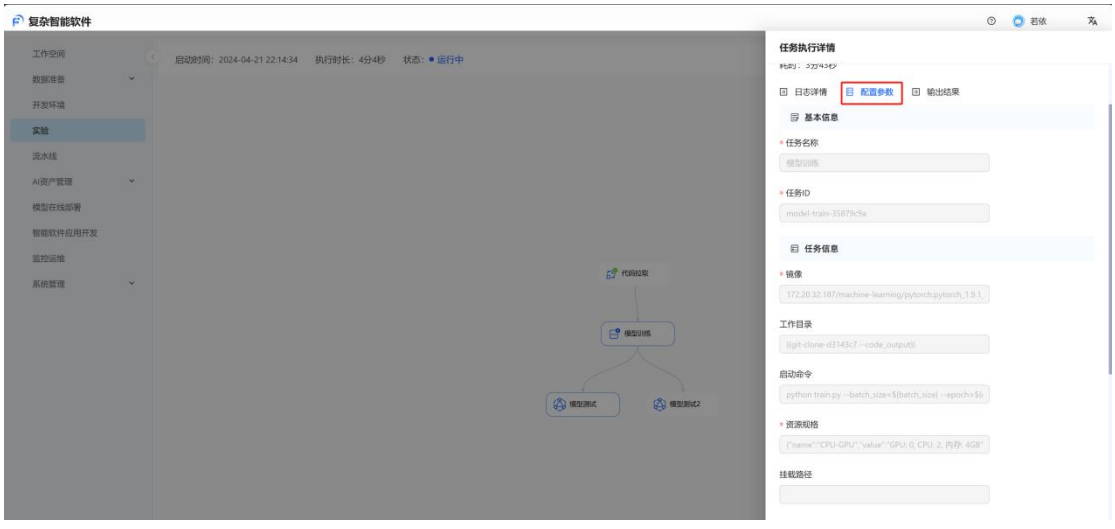
点击“执行参数”，可以查看此次实验的执行参数配置情况



双击实验任务，可以弹出任务的执行详情，包括运行状态，运行时间、日志详情。



在任务详情页点击“配置参数”可以查看当前任务的配置参数



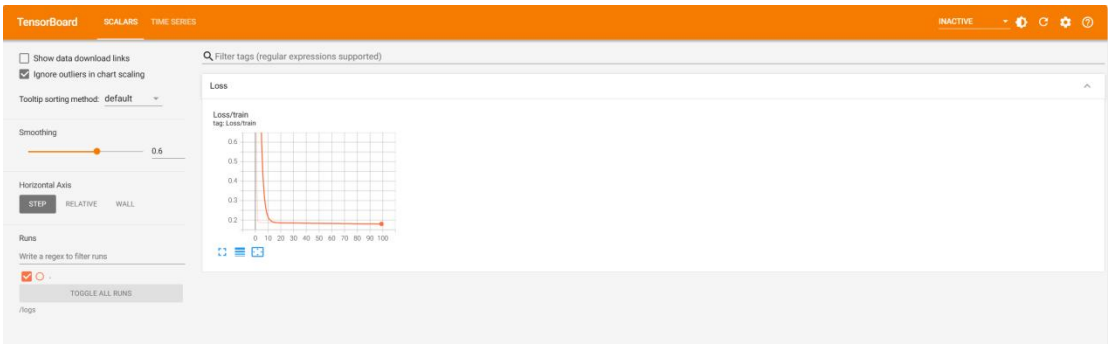
任务运行结束，点击“输出结果”可以查看和下载组件的输出结果。

4.4 实验过程参数变化查看

对于模型训练，有时候需要分析和查看模型训练过程中参数的变化，平台集成了 tensorboard 可视化面板查看参数的变化。模型训练和分布式训练组件默认开启 tensorboard 可视化记录（用户将 tensorboard 的 log 文件写入/tesorboard-logs 目录即可）

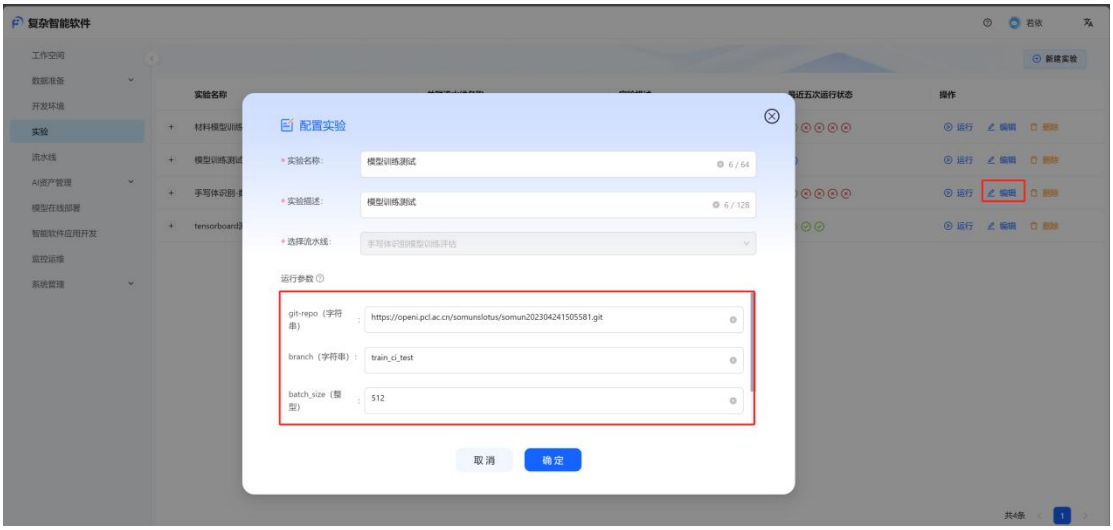
点击绿色启动按钮，即可启动 tensorboard 可视化

启动成功后，点击链接图标，即可查看 tensorboard 面板



4.5 实验动态传参

在实验界面，点击“编辑”，在弹出的页面中，可以修改实验参数。



每次实验的时候，修改实验参数，然后点击运行，这样就可以对一个实验进行多次运行，每次运行的参数不同，后续会上线指标对标的功能，可以方便的对比不同参数情况下指标的变化。

1. 删除实验

实验如果已经有运行的实例，不能直接删除，需要删除实例后才能删除实验，如果实验没有实例，可以直接删除。

5. AI 资产管理

5.1 数据集管理

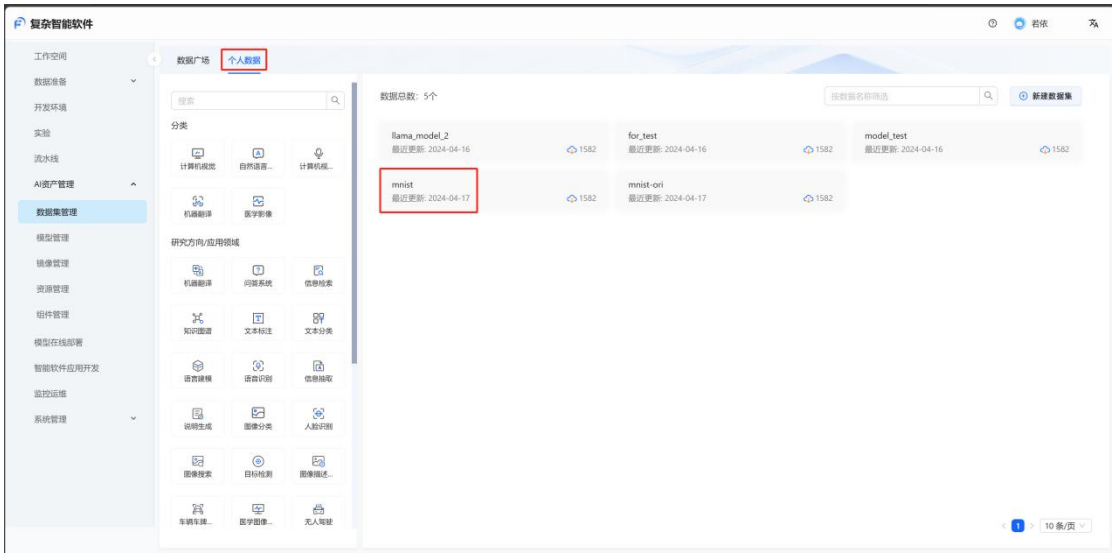
数据集管理包含数据集集广场和个人数据集。数据集广场汇聚了深度学习的常用的数据集。个人数据集则是用户上传的数据集。

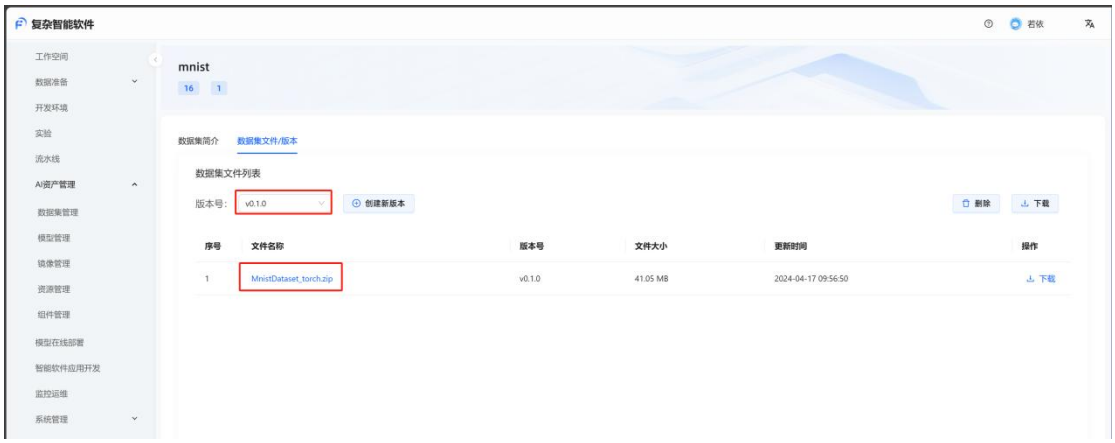
5.1.1 新建数据集

点击数据集管理，切换到“个人数据”，点击“新建数据集”，弹出新建数据集页面，填写好必填的信息即可新建数据集。

5.1.2 查看数据集

在“个人数据”标签页，点击数据集名称，既可以查看数据集的简介、版本、文件。



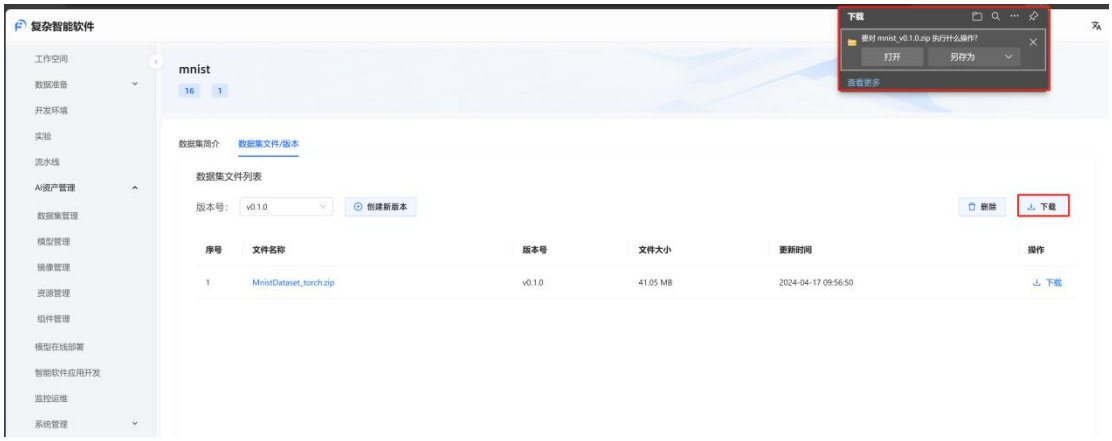


5.1.3 新增数据集版本

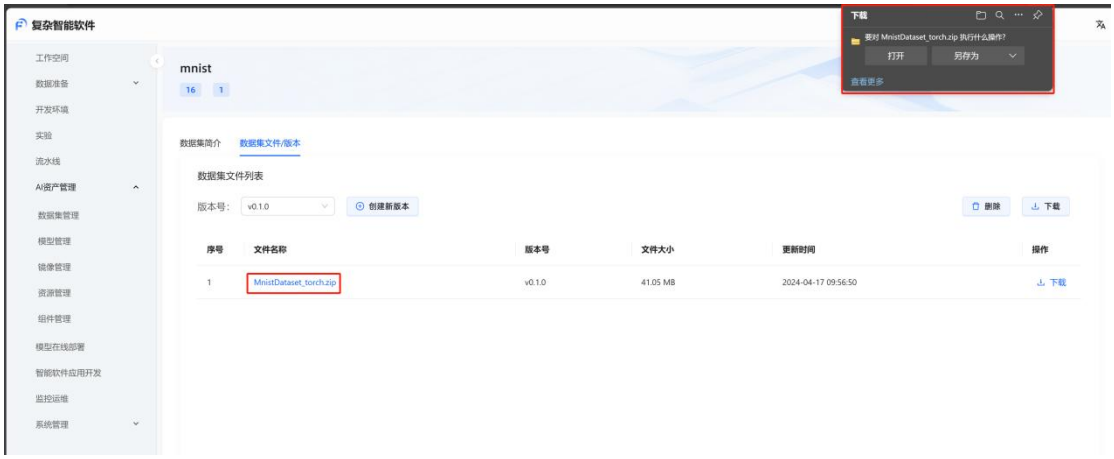
在数据集详情页面，点击“创建新版本”，填写必要的参数，即可新增数据集版本。

5.1.4 下载数据集

在数据集详情页面，点击“下载”，即可下载整个版本的数据集。



在数据集详情页面， 点击数据集的某个文件，即可下载该数据集文件。



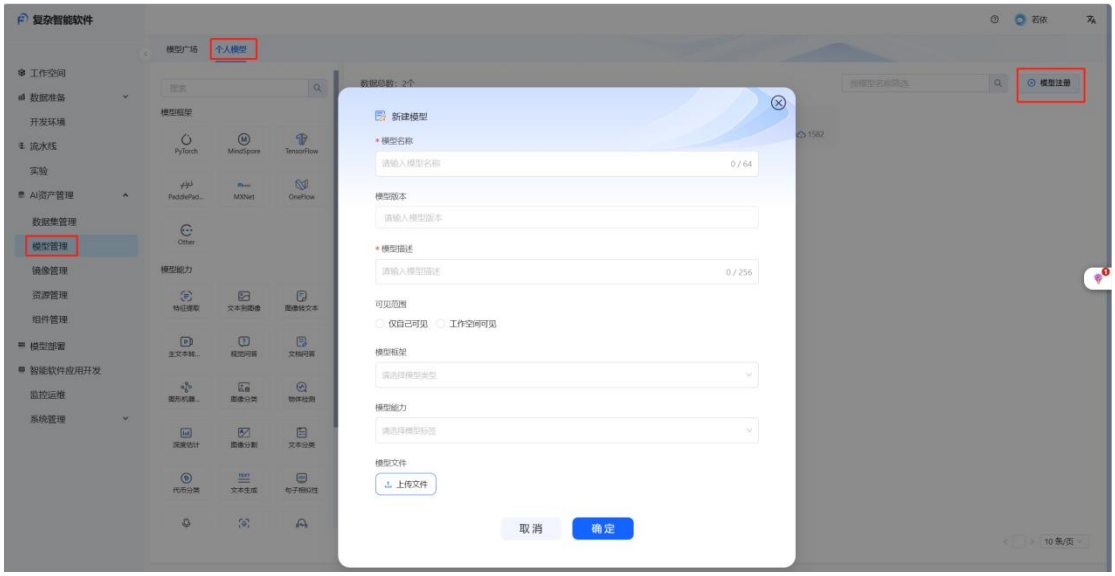
5.2 模型管理

模型管理包含模型广场和个人模型，模型广场汇聚了常用的经典的开源模型，个人模型则是

用户自己上传的模型。

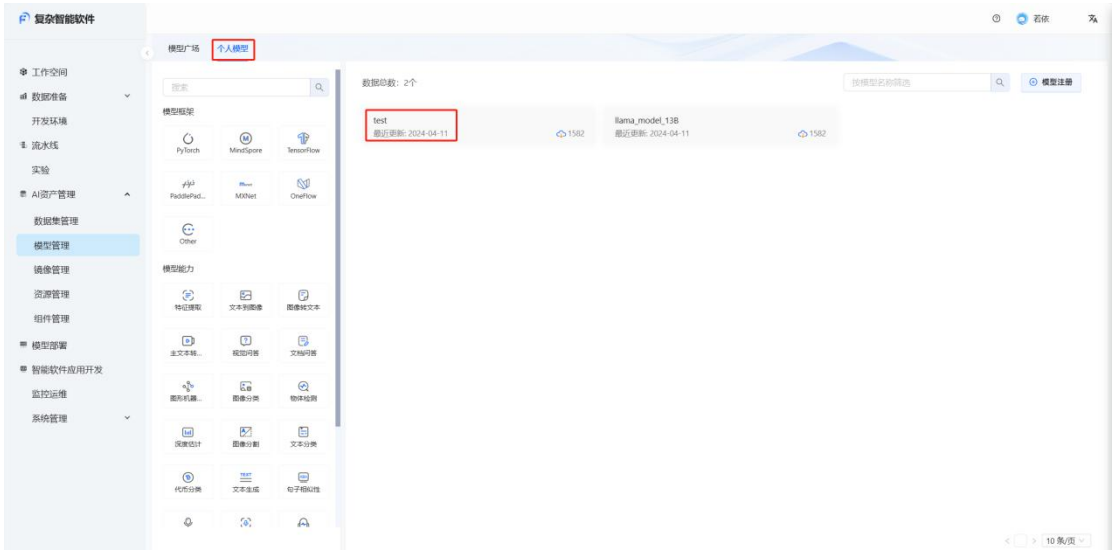
5.2.1 模型注册

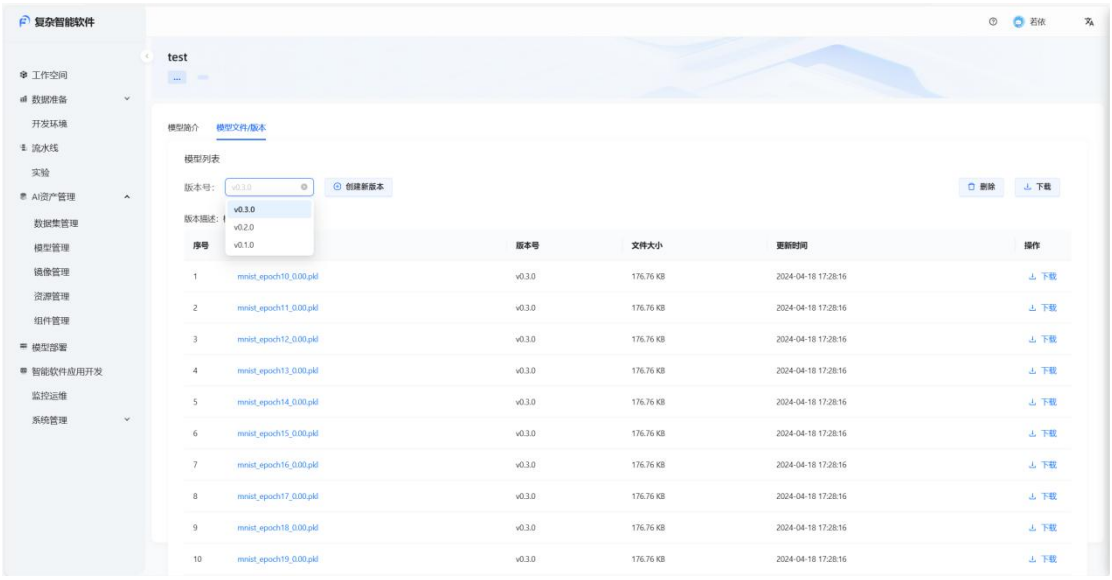
点击“模型管理”菜单，切换到“个人模型”，点击“模型注册”，在弹出的框中填写模型注册的必要信息，上传模型文件，即可注册模型。



5.2.2 模型详情查看

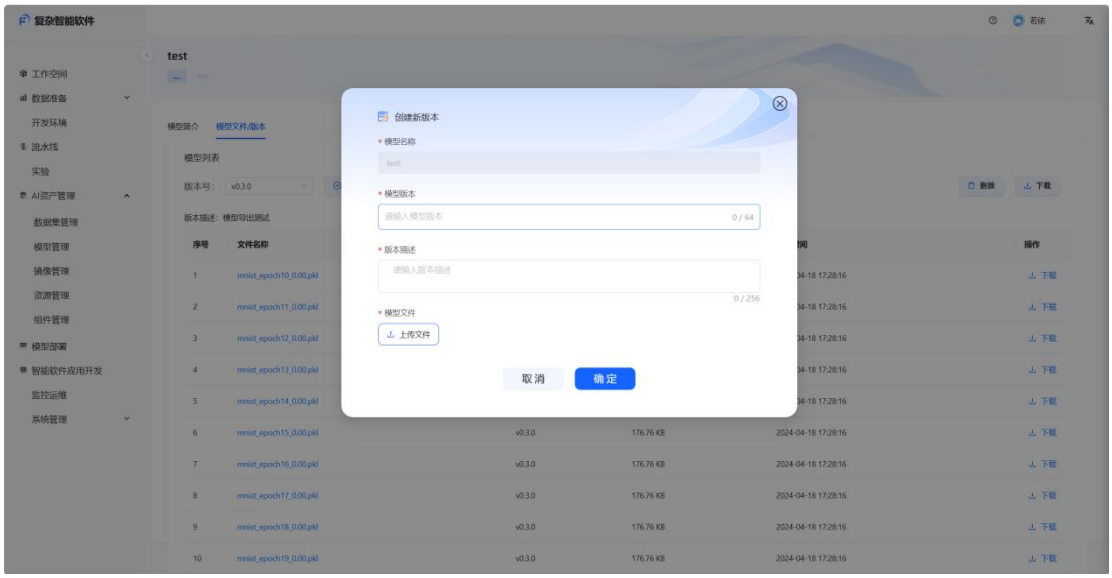
在“个人模型”标签页，点击模型的名称，可以进入模型详情页，查看模型的简介、版本信息、模型文件信息。





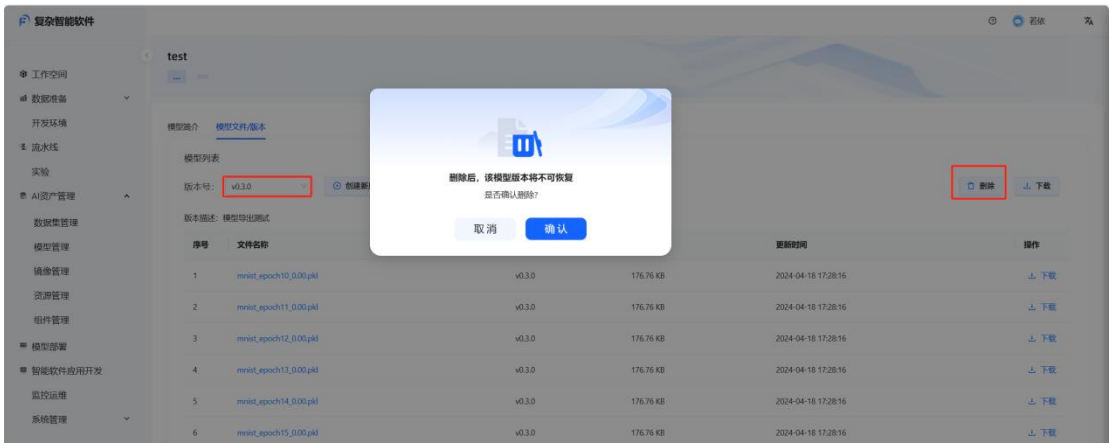
5.2.3 添加模型版本

在模型详情页面，点击“创建新版本”，即可创建新的模型版本。



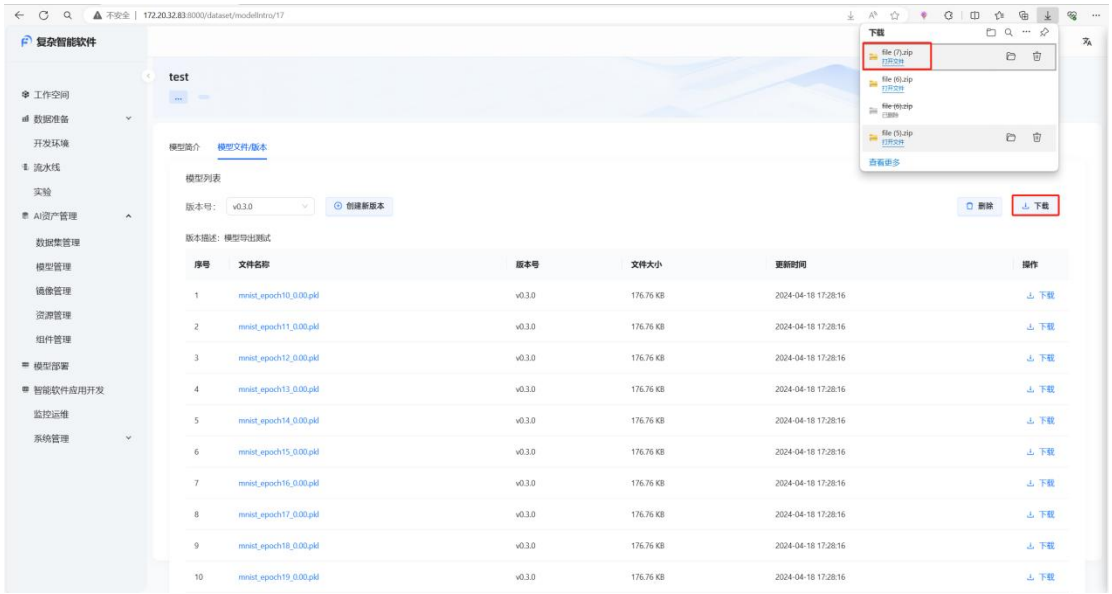
5.2.4 删除模型版本

在模型详情页面，点击“删除”，可以删除对应的模型版本，模型文件不能单独删除，只能删除整个模型版本。



5.2.5 下载模型

在模型详情页面，点击“下载”，即可下载选中的模型版本，模型版本以 zip 包的方式下载。



点击模型文件的某一个文件，可以对模型文件单独下载。

5.3 镜像管理

镜像管理提供了组件运行需要的镜像和各深度学习框架的镜像，同时提供制作镜像的功能，用户可以上传自己的镜像到镜像仓库中。

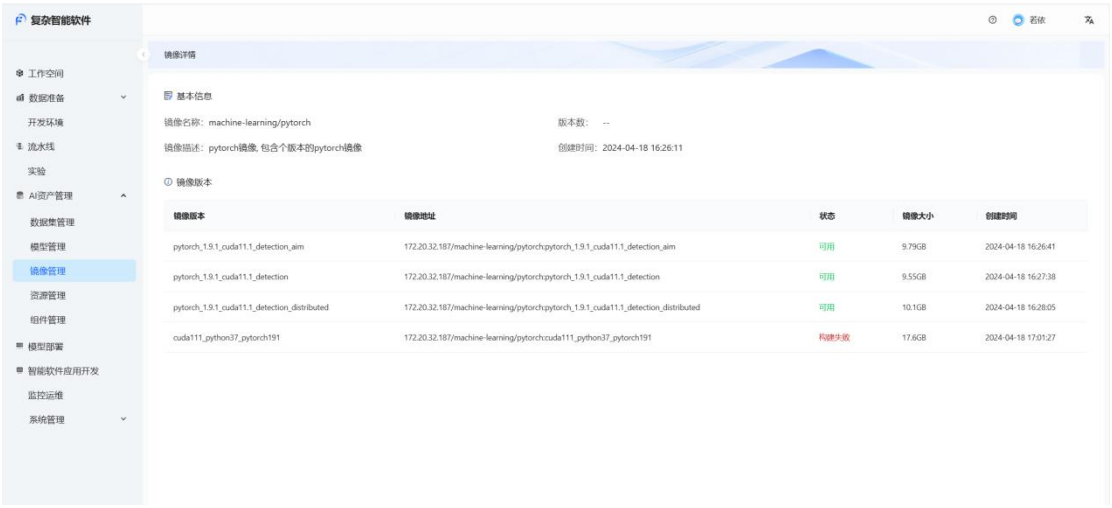
5.3.1 镜像列表查看

在菜单栏中点击“镜像管理”，可以查看公共镜像和个人镜像，默认显示公共镜像，如果需

要查看个人镜像，需要切换到“个人镜像”。

5.3.2 镜像版本查看

在镜像列表中，点击查看详情，可以查看该镜像的所有版本信息。



5.3.3 制作镜像

点击“镜像管理”菜单，切换到“个人镜像”，点击“制作镜像”，弹出镜像制作的文本框，镜像制作可以通过上传 tar 包方式和从公网镜像仓库导入两种方式。