ModelArts

数据准备

文档版本	01
发布日期	2026-04-16

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 数据准备功能说明

功能介绍

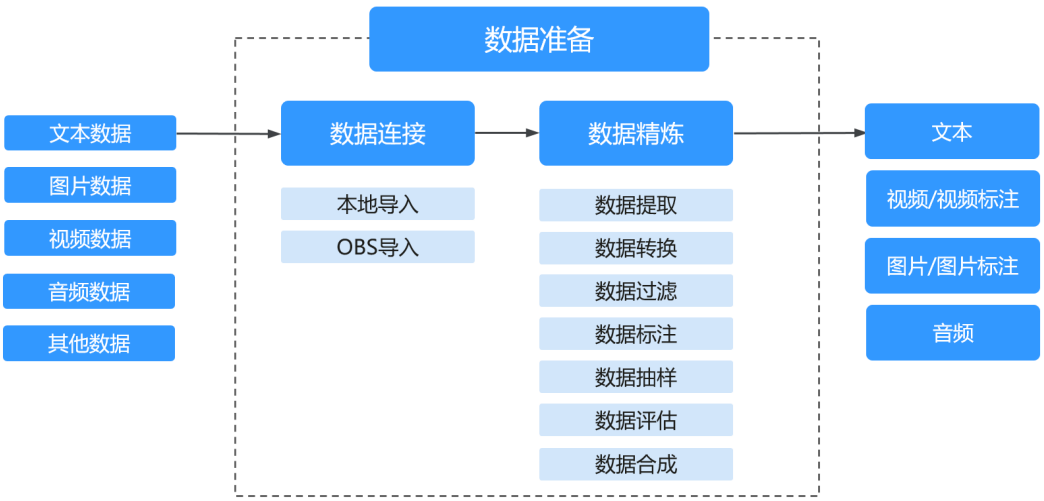
数据决定了大模型的能力上限。ModelArts**数据准备**功能提供了一站式、全流程的数据处理和管理服务，致力于解决大模型开发中“数据获取难、质量参差不齐、处理效率低”的核心痛点。通过内置的行业级数据处理算子与自动化流水线，系统化的处理数据获取、加工、发布等过程，帮助您将海量、多模态的原始数据，高效转化为高可用、高纯度的训练数据集，提高数据质量和处理效率，显著降低模型训练成本，提升模型泛化能力。

数据准备开发流程

ModelArts平台提供了全流程的数据开发功能，您可使用**数据连接**和**数据精炼**完成模型数据集的开发。其中**模型精炼**包含数据处理的全流程，包括数据加工，数据合成功能。助力开发者快速生成模型开发所需的数据集。

数据准备整体开发流程如图1-1所示。

图 1-1 数据准备开发流程



- **数据连接**：数据获取是数据工程的第一步，支持将不同来源和不同格式的数据导入平台，并生成“原始数据集”。通过该功能，用户可以轻松将大量数据导入平台，为后续的数据精炼和模型开发做好准备。详见**数据连接**章节。

- **数据精炼**：数据精炼模块提供了数据加工、数据合成一站式操作，旨在确保原始数据能够满足各种业务需求和模型开发的标准，加工出满足模型开发的数据集，详见**数据精炼**章节。

数据资产管理

数据资产管理模块为开发者提供了一站式的多种模态数据管理中心。它打破了数据孤岛，实现了从数据接入、版本控制、质量预览到最终调用的全链路闭环管理。ModelArts平台支持管理文本、图像、音频、视频等多种模态的数据，并根据来源不同，划分为平台预置数据资产与用户自定义数据资产两大类，满足从通用能力构建到垂直领域定制的全场景需求。详见**数据资产管理**章节。

ModelArts 平台支持的数据类型

ModelArts平台提供了业界最全面的数据处理功能。包括对文本类、图片类、音频类、视频类、平台格式数据集处理，还提供了自定义数据集功能，支持业界广泛使用的**Alpaca**和**ShareGPT**等数据集格式，能够灵活处理多样化的数据。

平台多样化的数据精炼和管理能力，为您提供丰富而全面的数据集，是您开发大模型的利器。

平台支持的数据类型见**表1-1**，各类型数据格式详细要求请参考**数据集格式要求**。

表 1-1 平台支持的数据类型

数据类型	数据内容	支持的文件格式	数据集要求
文本	文档	docx、pdf。	**文本类数据集格式要求**
	预训练文本	jsonl	
	单轮问答	jsonl、csv	
	单轮问答（人设）	jsonl、csv	
	多轮问答	jsonl	
	多轮问答（人设）	jsonl	
	问答排序	jsonl、csv	
	偏好优化DPO	jsonl	
	偏好优化DPO（人设）	jsonl	

数据类型	数据内容	支持的文件格式	数据集要求
图片类	图片	• 图片+jsonl（可选） – 图片格式支持：jpg、jpeg、png、bmp。 – jsonl为非必选文件类型。当存在jsonl文件时，需要保证如下条件： jsonl中索引的图片文件必须存在。 jsonl文件必须位于数据集根目录，且命名为annotation.jsonl。 jsonl文件仅支持UTF-8编码。	图片类数据集格式要求
视频类	视频	mp4、avi	视频类数据集格式要求
	视频+标注	• 视频+jsonl – 视频格式支持：mp4、avi。 – 标注文件格式：jsonl，jsonl文件仅支持UTF-8编码。	
音频类	音频	• 音频+jsonl – 音频文件：支持mp3、flac、wav、opus、aac、m4a格式，允许放在根目录或下层目录中。 – 标注文件格式：可选，格式为UTF-8编码的jsonl文件，每一行描述一个音频文件在数据集中的相对路径以及其它信息。	音频类数据集格式要求
其他类	自定义	支持构建用户自定义场景下所需的数据集类型。支持主流Alpaca和ShareGPT格式数据集。	其他类数据集格式要求

2 数据集格式要求

2.1 文本类数据集格式要求

ModelArts支持创建文本类数据集，创建时可导入多种形式的数据，具体格式要求详见表2-1。

约束限制

- 仅西南-贵阳一区域的新版控制台支持。
- 从OBS导入：单个文件/压缩包大小不超过20GB；多个文件场景，文件数量不限制，总文件大小不超过20GB。
- 本地导入：单个文件大小不超过1GB，文件数量最多20个。
- jsonl文件格式仅支持UTF-8编码。

表 2-1 文本类数据集格式要求

文件内容	文件格式	格式说明
文档	docx、pdf	原始文档内容。

预训练文本	jsonl	text表示预训练所使用的文本数据，具体格式示例如下：
		{
		"text": "《活着》，是中国著名作家余华所写的一部长篇小说。《活着》讲述了一个普通农民徐福贵的人生历程。他的人生充满了苦难和挫折，但他在面对这些困难时，始终保持着坚强和乐观的态度。"
		}

文件内容	文件格式	格式说明

单轮问答	jsonl	数据为jsonl格式时，支持**Alpaca/ShareGPT/标准格式**，以下为不同格式数据集示例。
		• **Alpaca格式**
		非思维链数据：数据由问答对构成，instruction、output分别表示问题、答案，具体格式示例如下：
		`{ "instruction": "请推荐一本书。", "input": "", "output": "当然可以，我推荐你阅读《自动驾驶的未来》。" }`
		思维链数据：数据由问答对构成，instruction、output分别表示问题、答案，并且output必须包含think标签对表示思考过程，具体格式示例如下：
		`{ "instruction": "能给我推荐点书吗？", "input": "", "output": "<think>用户请求图书推荐但未提供具体偏好，适合选择一本覆盖面广且具有前瞻性的科技书籍。</think>推荐《自动驾驶的未来》。" }`
		• **ShareGPT格式**
		非思维链数据：数据由问答对构成，对话中来自human、gpt角色的值分别表示问题、答案，具体格式示例如下：
		`{ "conversations": [ { "from": "human", "value": "能给我推荐点书吗？" }, { "from": "gpt", "value": "作为书籍推荐专家，我推荐你阅读《自动驾驶的未来》。" } ] }`
		思维链数据：数据由问答对构成，对话中来自human、gpt角色的值分别表示问题、答案，并且gpt角色的值必须包含think标签对表示思考过程，具体格式示例如下：
		`{ "conversations": [ { "from": "human", "value": "能给我推荐点书吗？" }, { "from": "gpt", "value": "<think>作为书籍推荐专家，应基于当前科技趋势和大众阅读接受度进行推荐。</think>我推荐你阅读《自动驾驶的未来》。" } ] }`
		• **标准格式**
		非思维链数据：数据由问答对构成，context、target分别表示问题、答案，具体格式示例如下：
		`{ "context": "你好，请介绍自己", "target": "我是盘古大模型" }`

文件内容	文件格式	格式说明

		思维链数据：数据由问答对构成，context、target分别表示问题、答案，并且target必须包含think标签对表示思考过程，具体格式示例如下：
		{
		"context": "你好，请介绍自己",
		"target": "<think> 好的，用户让我介绍一下自己。首先，我需要明确用户的身份和使用场景……</think>我是盘古大模型"
		}

	csv	● 非思维链数据：csv文件的第一列对应context，第二列对应target，具体格式示例如下： "你好，请介绍自己","我是盘古大模型" ● 思维链数据：csv文件的第一列对应context，第二列对应target，并且target必须包含think标签对，具体格式示例如下： "你好，请介绍自己","<think>用户让我介绍一下自己。首先，我需要明确用户的身份和使用场景.....</think>我是盘古大模型"

<table>
<tr><th>文件内容</th><th>文件格式</th><th>格式说明</th></tr>
<tr><td>单轮问答（带人设）</td><td>jsonl</td><td>数据为jsonl格式时，支持Alpaca/ShareGPT/标准格式，以下为不同格式数据集示例。

Alpaca格式
非思维链数据：system表示人设，instruction、output分别表示问题、答案。具体格式示例如下：
<pre>{
  "system": "你是一名书籍推荐专家",
  "instruction": "请根据用户需求推荐书籍",
  "input": "",
  "output": "作为书籍推荐专家，我推荐你阅读《自动驾驶的未来》。"
}</pre>
思维链数据：system表示人设，instruction、output分别表示问题、答案， 并且output必须包含think标签对表示思考过程，具体格式示例如下：
<pre>{
  "system": "你是一名书籍推荐专家",
  "instruction": "请根据用户需求推荐书籍。",
  "input": "",
  "output": "<think>作为书籍推荐专家，应结合科技发展趋势与大众阅读接受度进行判断。</think>我推荐你阅读《自动驾驶的未来》。"
}</pre>

ShareGPT格式
非思维链数据：数据由问答对构成，system_prompt表示人设，对话中来自human、gpt角色对应的值分别表示问题、答案，具体格式示例如下：
<pre>{
  "system_prompt": "角色名称: 书籍推荐专家。",
  "conversations": [
    {
      "from": "human",
      "value": "能给我推荐点书吗？"
    },
    {
      "from": "gpt",
      "value": "作为书籍推荐专家，我推荐你阅读《自动驾驶的未来》。"
    }
  ]
}</pre>
思维链数据：数据由问答对构成，system_prompt表示人设，对话中来自human、gpt角色对应的值分别表示问题、答案，并且gpt角色的值必须包含think标签对表示思考过程，具体格式示例如下：
<pre>{
  "system_prompt": "角色名称: 书籍推荐专家。",
  "conversations": [
    {
      "from": "human",
      "value": "能给我推荐点书吗？"
    },
    {
      "from": "gpt",
      "value": "<think>作为书籍推荐专家，应基于当前科技趋势和大众阅读接受度进行推荐。</think>我推荐你阅读《自动驾驶的未来》。"
    }
  ]
}</pre>

标准格式

</td></tr>
</table>

文件内容	文件格式	格式说明

		非思维链数据：数据由问答对构成，system表示人设，context、target分别表示问题、答案，具体格式示例如下：
		{
		"system": "机智幽默",
		"context": "你好，请介绍自己",
		"target": "哈哈，你好呀，我是你的聪明助手。"
		}
		思维链数据：数据由问答对构成，system表示人设，context、target分别表示问题、答案，并且target必须包含think标签对表示思考过程，具体格式示例如下：
		{
		"system": "机智幽默",
		"context": "你好，请介绍自己",
		"target": "<think> 好的，用户让我介绍一下自己。首先，我需要明确用户的身份和使用场景……</think>我是盘古大模型"
		}

	csv	● csv非思维链数据：csv文件的第一列对应system，第二三列分别对应context、target。具体格式示例如下： "你是一个机智幽默问答助手","你好，请介绍自己","哈哈，你好呀，我是你的聪明助手。" ● csv思维链数据：csv文件的第一列对应system，第二三列分别对应context、target， 并且target必须包含think标签对表示思考过程，具体格式示例如下： "你是一个机智幽默问答助手","你好，请介绍自己","<think>用户让我介绍一下自己。首先，我需要明确用户的身份和使用场景</think>哈哈，你好呀，我是你的聪明助手。"

<table>
<tr><th>文件内容</th><th>文件格式</th><th>格式说明</th></tr>
<tr><td>多轮问答</td><td>jsonl</td><td>数据为jsonl格式时，支持Alpaca/ShareGPT/标准格式，以下为不同格式数据集示例。

Alpaca格式
非思维链数据：数组格式，由一轮或多轮问答对构成。instruction、output分别表示问题、答案，history表示历史对话。具体格式示例如下：
<pre>{
  "instruction": "我想找自动驾驶相关的书籍。",
  "input": "",
  "output": "我推荐你阅读《自动驾驶的未来》。",
  "history": [
    [
      "我想找一本能了解未来科技趋势的书。",
      "未来科技涵盖面很广，你更关注哪些方向？"
    ]
  ]
}</pre>
思维链数据：数组格式，由一轮或多轮问答对构成，其中instruction、output分别表示问题、答案，history表示历史对话。并且至问答output包含think标签对表示思考过程，具体格式示例如下：
<pre>{
  "instruction": "我想找自动驾驶相关的书籍。",
  "input": "",
  "output": "<think>用户已明确具体领域，应推荐一本结构完整、兼顾技术与产业视角的代表性书籍。</think>我推荐你阅读《自动驾驶的未来》。",
  "history": [
    [
      "我想找一本能了解未来科技趋势的书。",
      "<think>作为书籍推荐专家，第一步应确认用户关注的具体技术领域，而非直接给出书名。</think>未来科技涵盖面很广，你更关注哪些方向？"
    ]
  ]
}</pre>
ShareGPT格式
非思维链数据：数组格式，由一轮或多轮问答对构成。对话中来自human、gpt角色对应的值分别表示问题、答案，具体格式示例如下：
<pre>{
  "conversations": [
    {
      "from": "human",
      "value": "你好"
    },
    {
      "from": "gpt",
      "value": "嗨！你好，需要点什么帮助吗？"
    },
    {
      "from": "human",
      "value": "能给我推荐点书吗？"
    },
    {
      "from": "gpt",
      "value": "当然可以，基于你的兴趣，我推荐你阅读《自动驾驶的未来》。"
    }</pre>
</td></tr>
</table>

文件内容	文件格式	格式说明

		]
		}
		思维链数据：数组格式，由一轮或多轮问答对构成，其中对话中来自human、gpt对应的值分别表示问题、答案，并且gpt角色的值包含think标签对表示思考过程，具体格式示例如下：
		{
		"conversations": [
		{
		"from": "human",
		"value": "我想了解未来科技趋势的书。"
		},
		{
		"from": "gpt",
		"value": "<think>先确认具体领域。</think>未来科技很广，你关注哪些？"
		},
		{
		"from": "human",
		"value": "自动驾驶。"
		},
		{
		"from": "gpt",
		"value": "<think>推荐代表性书籍。</think>我推荐《自动驾驶的未来》。"
		}
		]
		}
		• **标准格式**
		非思维链数据：数组格式，由一轮或多轮问答对构成。context、target分别表示问题、答案，具体格式示例如下：
		[
		{
		"context": "你好",
		"target": "你好，请问有什么可以帮助你的？"
		},
		{
		"context": "请介绍一下华为云的产品。",
		"target": "华为云提供包括但不限于计算、存储、网络等产品服务。"
		}
		]
		思维链数据：数组格式，由一轮或多轮问答对构成，其中context、target分别表示问题、答案，并且至少有一轮问答的target包含think标签对表示思考过程，具体格式示例如下：
		[
		{
		"context": "你好",
		"target": "你好，请问有什么可以帮助你的？"
		},
		{
		"context": "请介绍一下华为云的产品。",
		"target": "<think> 好的，用户让我介绍一下华为云产品。首先，我需要回忆一下……</think>华为云提供包括但不限于计算、存储、网络等产品服务。"
		}
		]

<table>
<tr><th>文件内容</th><th>文件格式</th><th>格式说明</th></tr>
<tr><td>多轮问答（带人设）</td><td>jsonl</td><td>数据为jsonl格式时，支持Alpaca/ShareGPT/标准格式，以下为不同格式数据集示例。

Alpaca格式
非思维链数据：数组格式，由一轮或多轮问答对构成。system表示人设，instruction、output分别表示问题、答案，history表示历史对话。具体格式示例如下：
<pre>{
  "system": "你是一名书籍推荐专家",
  "instruction": "我想找自动驾驶相关的书籍。",
  "input": "",
  "output": "我推荐你阅读《自动驾驶的未来》。",
  "history": [
    [
      "我想找一本能了解未来科技趋势的书。",
      "未来科技涵盖面很广，你更关注哪些方向？"
    ]
  ]
}</pre>
思维链数据：数组格式，由一轮或多轮问答对构成，其中system表示人设，instruction、output分别表示问题、答案，history表示历史对话。并且问答output包含think标签对表示思考过程，具体格式示例如下：
<pre>{
  "system": "你是一名书籍推荐专家",
  "instruction": "我想找自动驾驶相关的书籍。",
  "input": "",
  "output": "<think>用户已明确具体领域，应推荐一本结构完整、兼顾技术与产业视角的代表性书籍。</think>我推荐你阅读《自动驾驶的未来》。",
  "history": [
    [
      "我想找一本能了解未来科技趋势的书。",
      "<think>作为书籍推荐专家，第一步应确认用户关注的具体技术领域，而非直接给出书名。</think>未来科技涵盖面很广，你更关注哪些方向？"
    ]
  ]
}</pre>
ShareGPT格式
非思维链数据：数组格式，由一轮或多轮问答对构成。system_prompt表示人设，对话中来自human、gpt角色对应的值分别表示问题、答案，具体格式示例如下：
<pre>{
  "system_prompt": "书籍推荐专家",
  "conversations": [
    {
      "from": "human",
      "value": "你好"
    },
    {
      "from": "gpt",
      "value": "你好，需要什么帮助？"
    },
    {
      "from": "human",
      "value": "推荐点书？"
    },
    {
      "from": "gpt",
      "value": "我推荐《自动驾驶的未来》。"</pre>
</td></tr>
</table>

文件内容	文件格式	格式说明

		<pre> }
		  ]
		}</pre>
		思维链数据：数组格式，由一轮或多轮问答对构成，其中system_prompt表示人设，对话中来自human、gpt角色对应的值分别表示问题、答案，并且gpt角色的值包含think标签对表示思考过程，具体格式示例如下：
		<pre>{
		  "system_prompt": "书籍推荐专家",
		  "conversations": [
		    {
		      "from": "human",
		      "value": "我想了解未来科技趋势的书。"
		    },
		    {
		      "from": "gpt",
		      "value": "<think>先确认具体领域。</think>未来科技很广，你关注哪些方向？"
		    },
		    {
		      "from": "human",
		      "value": "自动驾驶。"
		    },
		    {
		      "from": "gpt",
		      "value": "<think>推荐代表性书籍。</think>我推荐《自动驾驶的未来》。"
		    }
		  ]
		}</pre>
		• **标准格式**
		非思维链数据：数组格式，由一轮或多轮问答对构成。system表示人设，context、target分别表示问题、答案，具体格式示例如下：
		<pre>[
		  {
		    "system": "机智幽默"
		  },
		  {
		    "context": "你好，请介绍自己",
		    "target": "哈哈，你好呀，我是你的聪明助手。"
		  }
		]</pre>
		思维链数据：数组格式，由一轮或多轮问答对构成，其中system表示人设，context、target分别表示问题、答案，并且至少有一轮问答的target包含think标签对表示思考过程，具体格式示例如下：
		<pre>[
		  {
		    "system": "机智幽默"
		  },
		  {
		    "context": "你好，请介绍自己",
		    "target": "<think> 好的，用户让我介绍一下自己。首先，我需要明确用户的身份和使用场景……</think>哈哈，你好呀，我是你的聪明助手。"
		  }
		]</pre>

文件内容	文件格式	格式说明

问答排序	jsonl	context表示问题，targets答案1、2、3表示答案的优劣顺序，最好的答案排在最前面
		{
		"context": "context内容",
		"targets": [
		"a",
		"b",
		"c"
		]
		}

	csv	● csv格式：csv文件的第一列对应context，其余列为答案。 "对于“沉舟侧畔千帆过”有怎样的理解","这句诗的本意是描述沉舟侧畔，新的船只仍然会像千帆一样驶过，这句话蕴含着深刻的哲理，表达了新事物必将取代旧事物的自然规律。","这句话可以用来形容自然景观，表达大自然的恢宏和生命的活力。"

<table>
<tr><th>文件内容</th><th>文件格式</th><th>格式说明</th></tr>
<tr><td>偏好优化DPO</td><td>jsonl</td><td>

非思维链数据：context表示问题，target表示期望的正确答案，bad_target表示不符合预期的错误答案。具体格式示例如下：
单轮DPO
<pre>{
  "context": [
    "你好，请介绍自己"
  ],
  "target": "我是盘古大模型",
  "bad_target": "我不会回答"
}</pre>多轮DPO
<pre>{
  "context": [
    "你好，请介绍自己",
    "我是盘古大模型",
    "请介绍一下华为云的产品。"
  ],
  "target": "华为云提供包括但不限于计算、存储、网络等产品服务。",
  "bad_target": "我不会回答"
}</pre>
思维链数据：context表示问题，target表示期望的正确答案，bad_target表示不符合预期的错误答案，答案中至少有一个包含think标签对表示思考过程，具体格式示例如下：
单轮DPO
<pre>{
  "context": [
    "你好，请介绍自己"
  ],
  "target": "<think> 好的，用户让我介绍一下自己。首先，我需要明确用户的身份和使用场景……</think>我是盘古大模型",
  "bad_target": "<think> 好的，用户让我介绍一下自己。……</think>我不会回答"
}</pre>多轮DPO
<pre>{
  "context": [
    "你好，请介绍自己",
    "我是盘古大模型",
    "请介绍一下华为云的产品。"
  ],
  "target": "<think> 好的，用户让我介绍一下华为云产品。首先，我需要回忆一下……</think>华为云提供包括但不限于计算、存储、网络等产品服务。",
  "bad_target": "<think> 好的，用户让我介绍一下华为云产品。……</think>我不会回答"
}</pre>

</td></tr>
</table>

文件内容	文件格式	格式说明

偏好优化DPO（人设）	jsonl	• 非思维链数据：system表示人设，context表示问题，target表示期望的正确答案，bad_target表示不符合预期的错误答案。具体格式示例如下：
		单轮DPO带人设
		{
		"system": "你是一位机智幽默的问答助手",
		"context": [
		"你好，请介绍自己"
		],
		"target": "哈哈，你好呀，我是你的聪明助手，怎么帮到你？",
		"bad_target": "我不会回答"
		}
		多轮DPO带人设
		{
		"system": "你是一位机智幽默的问答助手",
		"context": [
		"你好，请介绍自己",
		"哈哈，你好呀，我是你的聪明助手，怎么帮到你？",
		"请介绍一下有哪些产品。"
		],
		"target": "我们产品种类繁多，不仅涵盖计算、存储和网络，还有更多选择哦！",
		"bad_target": "我不会回答"
		}
		• 思维链数据：system表示人设，context表示问题，target表示期望的正确答案，bad_target表示不符合预期的错误答案，正确答案中至少有一个包含think标签对表示思考过程，具体格式示例如下：
		单轮DPO带人设
		{
		"system": "你是一位机智幽默的问答助手",
		"context": [
		"你好，请介绍自己"
		],
		"target": "<think>用户让我介绍一下自己。首先，我需要明确用户的身份和使用场景</think>哈哈，你好呀，我是你的聪明助手，怎么帮到你？",
		"bad_target": "我不会回答"
		}
		多轮DPO带人设
		{
		"system": "你是一位机智幽默的问答助手",
		"context": [
		"你好，请介绍自己",
		"哈哈，你好呀，我是你的聪明助手，怎么帮到你？",
		"请介绍一下有哪些产品。"
		],
		"target": "<think>客户想要了解产品</think>我们产品种类繁多，不仅涵盖计算、存储和网络，还有更多选择哦！",
		"bad_target": "我不会回答"
		}

2.2 图片类数据集格式要求

ModelArts支持创建图片类数据集，具体格式要求详见**表2-2**。

约束限制

- 仅西南-贵阳一区域的新版控制台支持。
- 从OBS导入：单个文件/压缩包大小不超过20GB；多个文件场景，文件数量不限制，总文件大小不超过20GB。
- 本地导入：单个文件大小不超过1GB，文件数量最多20个。
- jsonl文件格式仅支持UTF-8编码。

表 2-2 图片类数据集格式要求

文件内容	文件格式	文件要求

图片	图片+jsonl	• 图片：支持jpg、jpeg、png、bmp类型。
		 
		
		• 根目录下可存在单个annotation.jsonl文件，file_name字段必选。
		{
		"file_name": "图片名称（abc.jpg）"
		}

2.3 视频类数据集格式要求

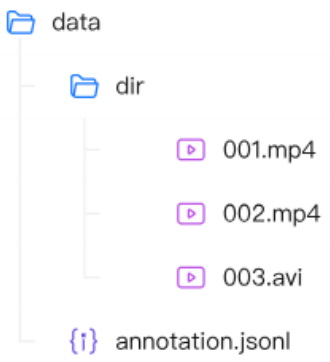
ModelArts支持创建视频类数据集，创建时可导入多种形式的数据，具体格式要求详见表2-3。

约束限制

- 仅西南-贵阳一区域的新版控制台支持。
- 从OBS导入：单个文件/压缩包大小不超过20GB；多个文件场景，文件数量不限制，总文件大小不超过20GB。
- 本地导入：单个文件大小不超过1GB，文件数量最多20个。
- jsonl文件格式仅支持UTF-8编码。

表 2-3 视频类数据集格式要求

文件内容	文件格式	文件要求
视频	mp4/avi	支持mp4、avi视频格式上传，所有视频可以放在多个文件夹下，每个文件夹下可以同时包含mp4或avi格式的视频。

视频+标注	视频+jsonl	• 视频格式支持：mp4、avi
		示例如下所示：
		 
		
		具体的jsonl标注文件参考：
		{
		"video_fn": "13/ad098173-af09-48fe-95c3-e72fd629688e.mp4",
		"prompt": "A person pours a clear liquid from a bottle into a shot glass, then lifts the glass to their mouth and drinks the shot. The background includes a red coat and other indistinct background elements.",
		"long_prompt": "A person is seen pouring a clear liquid from a green glass bottle into a small glass. The individual is wearing a white shirt with a lace collar and a beige cardigan. The background appears to be a cozy indoor setting, possibly a cafe or a restaurant, with red and white elements visible, such as a red coat hanging on the wall and a white table. The person carefully pours the liquid, ensuring it is filled to the brim of the glass. The liquid is clear and has some green leaves floating in it. The person then holds the glass up, possibly to show the contents or to prepare for a drink."
		}

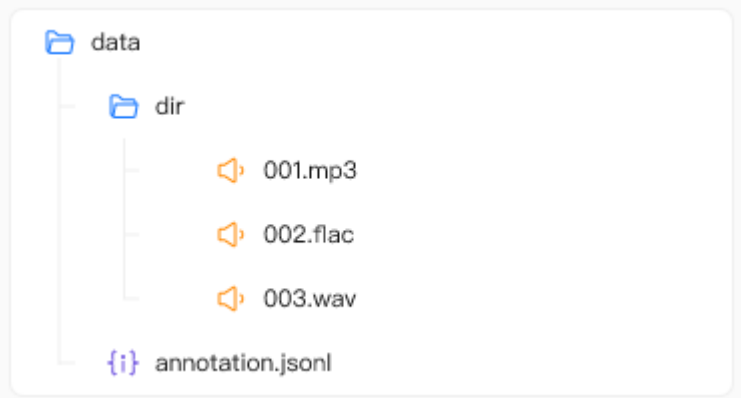
2.4 音频类数据集格式要求

ModelArts支持创建音频类数据集，具体格式要求详见表2-4。

约束限制

- 仅西南-贵阳一区域的新版控制台支持。
- 从OBS导入：单个文件/压缩包大小不超过20GB；多个文件场景，文件数量不限制，总文件大小不超过20GB。
- 本地导入：单个文件大小不超过1GB，文件数量最多20个。
- jsonl文件格式仅支持UTF-8编码。

表 2-4 音频类数据集格式要求

文件内容	文件格式	文件要求
音频	音频+jsonl（可选）	• 音频格式支持：mp3、flac、wav、opus、aac、m4a格式，允许放在根目录或下层目录中。 标注文件格式：可选，格式为UTF-8编码的jsonl文件，每一行描述一个音频文件在数据集中的相对路径以及其它信息。 示例如下所示：  具体的jsonl标注文件参考： {"audio_name":"dir/001.mp3","caption":"1"} {"audio_name":"dir/002.flac","caption":"2"} {"audio_name":"dir/003.wav","caption":"3"}

2.5 其他类数据集格式要求

除文本、图片、视频、音频数据集外，平台还支持导入其他类数据集，即用户训练模型时使用的自定义数据集，例如常用的开源Alpaca和ShareGPT格式数据集。

从OBS导入：单个文件/压缩包大小不超过20GB；多个文件场景，文件数量不限制，总文件大小不超过20GB。

本地上传：单个文件大小不超过1GB，单次上传文件数量最多20个。

本章将介绍常见的开源数据集格式要求。

Alpaca 数据集格式要求

Alpaca是开源模型（如DeepSeek系列、Qwen系列等）常用的数据集格式，是开源模型数据微调使用的主要数据集格式。特别用于instruction-tuning，即指令微调。其数据格式的特点是提供了一个明确的任务描述（instruction）、输入（input）和输出（output）三部分。

典型的Alpaca数据集格式：

```
[
  {
    "instruction": "人类指令（必填）",
```

```
    "input": "人类输入（选填）",
    "output": "模型回答（必填）",
    "system": "系统提示词（选填）",
    "history": [
      [
        "第一轮指令（选填）",
        "第一轮回答（选填）"
      ],
      [
        "第二轮指令（选填）",
        "第二轮回答（选填）"
      ]
    ]
  }
]
```

字段说明：

- instruction: 任务的指令，告诉模型需要完成什么操作。
- input: 任务所需的输入。如果任务是开放式的或者不需要明确的输入，这一字段可以为空字符串。
- output: 任务的期望输出，也就是模型在给定指令和输入情况下需要生成的内容。如果想训练带思考模式的模型，需要加<think></think>标签，或者引导思考的prompt，例如“Let's think step by step”。
- system：系统提示词（如什么风格、什么角色），该字段可选。
- history：是由多个字符串二元组构成的列表，分别代表历史消息中每轮对话的指令和回答。在指令监督微调时，历史消息中的回答内容也会被用于模型学习，该字段可选。

特点：

- Alpaca的数据格式结构简单易理解。
- 任务指令和输入内容是分离的，适合各种自然语言处理任务，如文本生成、翻译、总结等。

ShareGPT 数据集格式要求

ShareGPT格式来源于通过记录ChatGPT与用户对话的数据集，主要用于对话系统的训练。它更侧重于多轮对话数据的收集和组织，模拟用户与AI之间的交互。ShareGPT格式支持多种角色种类，例如human、gpt、observation、function等。它们按照不同角色对象在conversations列中呈现。

典型的ShareGPT数据集格式：

```
[
  {
    "conversations": [
      {
        "from": "human",
        "value": "人类指令"
      },
      {
        "from": "function_call",
        "value": "工具参数"
      },
      {
        "from": "observation",
        "value": "工具结果"
      },
      {
        "from": "gpt",
```

```
            "value": "模型回答"
        }
      ],
      "system":"系统提示词（选填）",
      "tools": "工具描述（选填）"
   }
]
```

- conversations：对话列表，包含每轮对话的角色及其对话内容，必选字段。其角色字段定义如下：
  - human：对话中人类发出的指令。
  - function_call：工具调用，这个工具就是一个AP，提供了某种功能。
  - observation：观测结果，即function_call的执行结果。
  - gpt：大模型根据人类下发指令的回答。

  注意：在角色中human和observation必须出现在奇数位置，gpt和function必须出现在偶数位置。
- system：系统提示词，可选字段。
- tools：工具，即对function_call的总结描述，可选字段。

特点：

ShareGPT格式更贴近人类与AI交互的方式，适用于构建和微调对话模型。

选择建议

- Alpaca格式适用于单轮指令微调，如任务型对话、问答系统或工具调用。其结构化设计简化了模型对明确指令的理解与响应，常用于轻量级微调（如LoRA）或基础能力训练（如文本生成、翻译）。
- ShareGPT格式专注于多轮对话场景，通过conversations字段记录用户与助手的交互历史，适合训练对话型模型（如聊天机器人、客服助手），尤其在上下文理解、情感对话或复杂推理等需要保持对话连贯性的任务中表现更优。
- 两者可结合使用，前者强化基础能力，后者提升交互体验。

3 数据连接

3.1 快速实现数据连接

ModelArts平台提供了方便的数据连接功能，您可以将自有数据集导入ModelArts后直接做训练模型。也可以通过**数据精炼**完成对数据集加工，加工出更多样化的数据集，对模型做更深入的开发。

本文将通过以下假设场景介绍如何使用ModelArts数据连接功能。

业务场景

ModelArts平台提供了最新的Qwen3大模型，您希望通过本地准备好的训练数据集对Qwen3模型做微调。

数据集为**Alpaca格式**，可以直接使用该数据集对Qwen3做微调。

您需要将本地数据集导入到ModelArts后做模型微调。

前提条件

1. 已注册华为账号并开通华为云，进行了实名认证，且在使用ModelArts前检查账号状态，账号不能处于欠费或冻结状态。
   - **注册华为账号并开通华为云**
   - **进行实名认证**
2. 配置委托访问授权

   ModelArts使用过程中涉及到OBS等服务交互，首次使用ModelArts需要用户配置委托授权，允许访问这些依赖服务。
3. 您本地要有能够训练Qwen3的训练数据集，数据集为Alpaca格式，格式说明参见**Alpaca数据集格式要求**。

约束限制

- 仅西南-贵阳一区域的新版控制台支持。

计费说明

数据连接计费涉及到数据存储OBS计费，具体可参考**数据管理计费项**。

步骤一：本地数据上传至 OBS

参考**OBS桶上传**操作上传数据。

步骤二：修改数据连接配置任务

1. 前往**ModelArts管理控制台**。
2. 在控制台左侧导航栏选择“数据准备 > 数据连接”，打开“数据连接”工作区，如**图3-1**所示。

图 3-1 数据连接工作区

3. 在“数据连接”工作区右上方单击“创建数据连接”按钮，打开“创建数据连接”配置页面。输入数据连接任务名称和描述。

图 3-2 “创建数据连接”配置任务

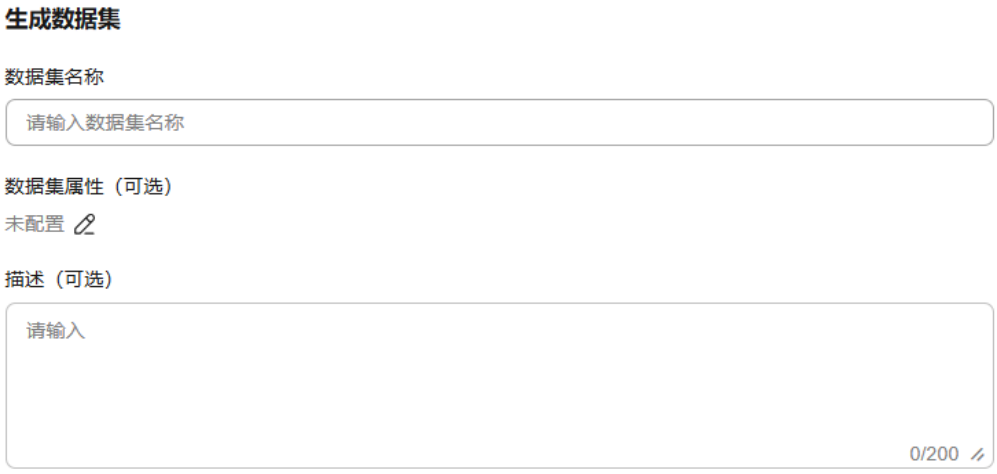
任务名称为必选，描述信息为可选，任务名称命名格式要求：以中文、字母**开头**，以中文、字母、数字**结尾**，长度2~64的字符。只允许输入中文、字母、数字、中划线、下划线等字符，具体参见**创建数据连接任务**中任务命名要求。

4. 导入数据集。在“数据导入”配置项选择数据集类型为“其他 > 自定义”。如**图3-3**所示。选择**连接方式**为**对象存储服务OBS**，将步骤一导入到OBS的数据作为本次数据集的来源。

图 3-3 数据导入

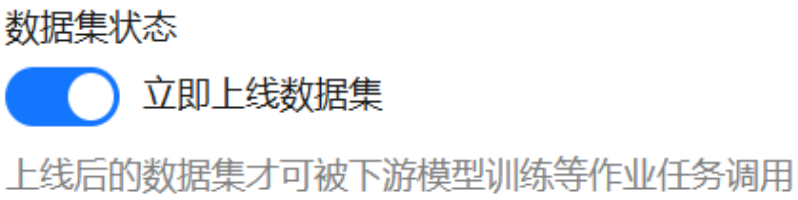
5. 将OBS导入数据作为一个数据集，需要给数据集重新命名。输入数据集名称、数据集属性（可选）、描述信息（可选），此时本地数据才算是ModelArts上的一个数据集。

图 3-4 填写生成数据集信息



6. 数据集填写完成后，配置“立即上线数据集”。
   - 选择立即上线数据集，生成数据集在“资产管理 > 数据 > 我的数据”为上线状态，可以被下游模型训练等作业直接使用。
   - 不选择立即上线数据集，生成数据集在“资产管理 > 数据 > 我的数据”为下线状态，不可被下游模型训练等作业直接使用，需要手动上线数据集后才能使用。

图 3-5 勾选“立即上线数据集”



7. 生成数据集还有一些扩展信息可以选填，说明数据集的版权信息，本文示例不填该信息。
8. 所有配置都已经完成，单击工作区右下角“立即创建”按钮，开始启动本次数据连接任务。待任务完成后，就可使用该数据集做Qwen3的微调工作了。
9. 连接任务完成后导入的数据集，可在控制台左侧选择“**资产管理 > 数据 >我的数据**”列表中查看。

相关参考

1. 数据集相关格式问题，请参见**数据集格式要求**。
2. 开源数据集说明请参见**其他类数据集格式要求**。

3.2 创建数据连接

使用场景

在数据处理和模型训练的场景中，用户需要将多种类型的数据集高效、准确地导入到ModelArts数据平台中，以支持后续的数据精炼和模型训练任务。然而，传统的数据导入方式存在诸多限制，如不支持自定义任务名称、数据格式转换功能有限等，导致用户在导入数据时面临操作不便和数据处理效率低下的问题。如何在新的平台中实现更加灵活和高效的数据导入功能，成为用户亟待解决的问题。为此，ModelArts平台提供了增强的数据导入功能，支持多种基础数据类型的导入，允许用户在创建导入任务时编辑任务名称和描述，同时支持数据格式转换和数据集的直接发布，从而显著提升了数据处理的灵活性和效率，满足了用户在数据准备阶段的多样化需求。

前提条件

1. 已注册华为账号并开通华为云，进行了实名认证，且在使用ModelArts前检查账号状态，账号不能处于欠费或冻结状态。
   - **注册华为账号并开通华为云**
   - **进行实名认证**
2. 配置委托访问授权

   ModelArts使用过程中涉及到OBS等服务交互，首次使用ModelArts需要用户配置委托授权，允许访问这些依赖服务。
3. 创建导入任务前，请先按照**数据集格式要求**提前准备数据。

约束限制

- 仅西南-贵阳一区域的新版控制台支持。
- 数据连接导入数据文件或压缩包不超过20GB，总文件大小也不超过20GB。

- 通过OBS导入数据，在指定OBS路径时只支持指定到文件夹，不支持指定到文件。

计费说明

数据连接计费涉及到数据存储OBS计费，具体可参考**数据管理计费项**。

创建数据连接任务

创建数据连接任务步骤如下：

1. 前往**ModelArts管理控制台**。
2. 在控制台左侧导航栏选择“数据准备 > 数据连接”，打开“数据连接”工作区，如**图3-6**所示。

图 3-6 数据连接

3. 在**数据连接**工作区右上方单击“创建数据连接”按钮，打开“创建数据连接”配置页面。输入数据连接任务名称和描述。任务名称为必选，描述信息为可选。

图 3-7 创建数据连接

说明：

任务名称：命名默认为data-connect-年月日时分秒，如：data-connect-20260116233300，也可自定义名称，命名要求如下：

- 命名长度：2~64字符。
- 格式要求：以中文、字母**开头**，以中文、字母、数字**结尾**。只允许输入中文、字母、数字、中划线、下划线等字符。

描述：无格式要求，长度不超过200字符，内容可选填。

4. 导入数据至ModelArts平台，如**图3-8**所示。请根据具体使用场景导入对应类型的数据，**连接方式**为**对象存储服务OBS**或**本地上传**。

单轮问答、单轮问答（带人设）、多轮问答、多轮问答（带人设）支持**Alpaca格式 /ShareGPT格式/平台格式**三种文件格式。

图 3-8 数据导入

5. 通过数据导入功能将数据导入到ModelArts平台，生成数据集。数据集可用于数据精炼、模型训练、微调、评测等用途。

   导入的数据形成新的数据集，需要给数据集重新命名。输入数据集名称、数据集属性（可选）、描述信息（可选）。

   说明：

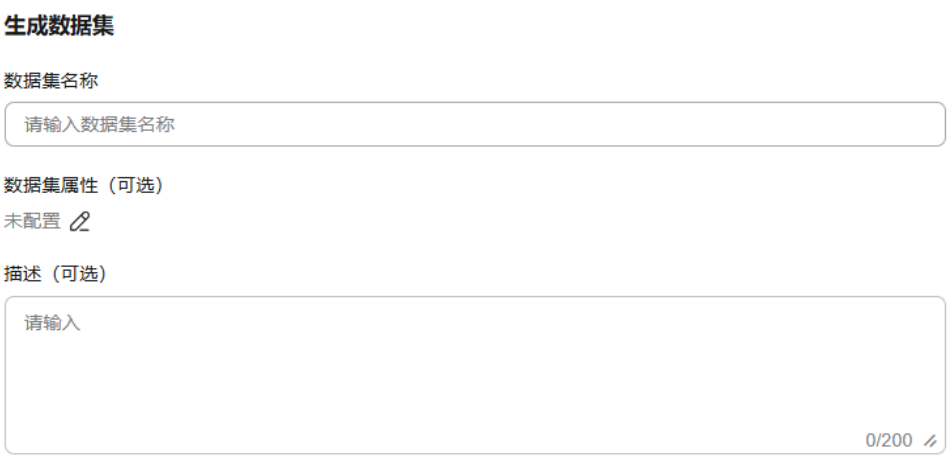
   数据集名称：

   - 命名长度：2~63字符。
   - 格式要求：以中文、字母**开头**，以中文、字母、数字**结尾**。只允许输入中文、字母、数字、中划线、下划线字符。

   数据集属性：可选字段，支持配置标签。可以按照行业、语言维度配置标签，也可自定义标签。

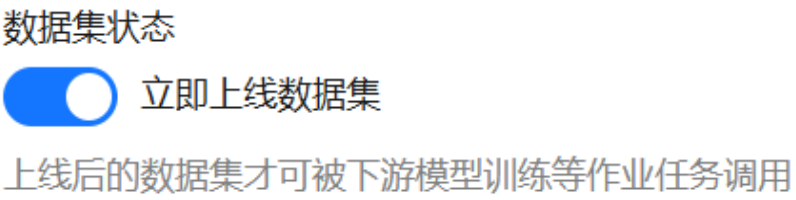
   描述：可选字段，无格式要求，长度不超过200字。

图 3-9 生成原始数据集



6. 数据集填写完成后，配置“立即上线数据集”。
   - 选择立即上线数据集，生成数据集在“**资产管理 > 数据 >我的数据**”为上线状态，可以被下游模型训练等作业直接使用。
   - 不选择立即上线数据集，生成数据集在“**资产管理 > 数据 >我的数据**”为下线状态，不可被下游模型训练等作业直接使用，需要手动上线数据集后才能使用。

   图 3-10 勾选“立即上线数据集”

7. 生成数据集扩展信息可以选填，扩展信息包括**数据集版权**相关信息，数据集版权功能主要用于记录和管理数据集的版权信息，确保数据的使用合法合规，明确数据集的来源和版权授权人、数据集许可证等信息。通过填写这些信息，可以追溯数据的来源，明确数据使用的限制和许可，从而保护数据版权并避免版权纠纷。

   图 3-11 扩展信息

8. 单击页面右下角“立即创建”，返回至“数据连接”页面，在该页面可以查看数据集的任务状态，如果状态为“运行成功”，则数据连接任务成功。

   图 3-12 查看数据连接任务列表

9. 数据集生成后，可在“**资产管理 > 数据 >我的数据**”列表查看生成数据集。

3.3 管理数据连接

约束限制

- 仅西南-贵阳一区域的新版控制台支持。

功能介绍

数据连接需要对所有任务做统一管理。在数据连接任务工作区可以查看数据连接任务的名称/ID、状态，生成数据集的名称、操作时间、操作人以及支持的一些操作。如图3-13所示。本文将详细介绍如何管理数据连接任务。

图 3-13 查看数据连接任务列表

展示/隐藏功能简介

打开数据连接工作区，最上方默认会展示数据连接的功能简介，用户能快速感知数据连接的功能亮点及业务范围，快速上手无难度。

图 3-14 打开功能简介

对于经验丰富的用户，也可单击功能简介，关闭该功能。如图3-15所示。

图 3-15 关闭功能简介

数据连接任务管理

数据连接任务支持过滤、搜索、删除、重试等操作。以下分别讲解如何操作。

- **数据连接任务过滤和搜索。**在数据任务列表比较多的情况下，支持按照**任务ID、任务名称、状态、生成数据集、操作人**等维度过滤想要的任务，便于快速找到目标任务。也可单击“我创建的”开关，只列举当前登录用户创建的数据连接任务。

图 3-16 数据集过滤

- **数据连接任务删除。**

  对于列表中创建的数据连接任务，支持“删除”操作。可以在“操作”列单击“删除”按钮，在弹出删除对话框选择“确定”后，该任务将被删除。删除后的任务不是彻底删除，为避免误删，如果还想再继续使用，可以恢复任务。已删除的任务在任务清单有“已删除”标签，如图3-18所示。对于已删除的任务，可以选择“彻底删除”，如图3-19所示。彻底删除后的任务不可恢复。

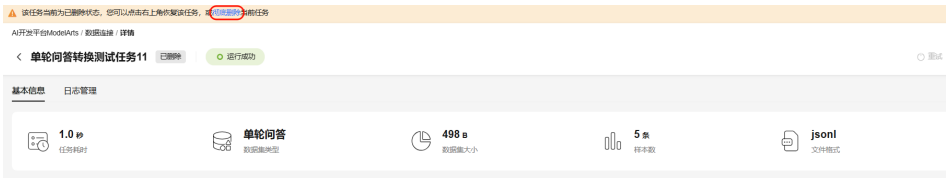
  图 3-17 删除任务

  图 3-18 已删除任务的“已删除”标签

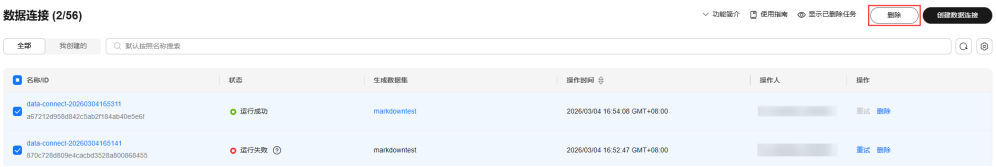
  图 3-19 任务“彻底删除”

- **数据连接任务批量删除。**

  勾选数据连接任务名称前的复选框，选择要删除的任务后，单击右上角“删除”按钮，可以批量删除数据。

  图 3-20 批量删除

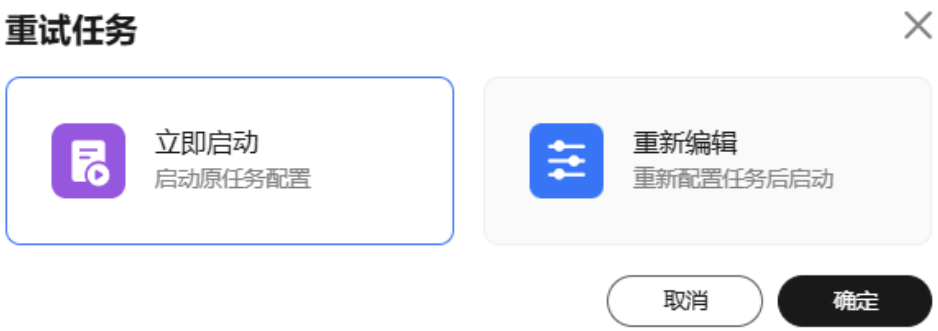
- **数据连接任务重试。**

  对于运行失败的任务，支持“重试”操作重新运行该任务。可以在“操作”列单击“重试”按钮，在弹出的对话框中选择“立即启动”或“重新编辑”，如图3-21所示。两种重试任务的区别如下：

  立即启动：不修改连接任务的任何配置，直接重新启动任务。

  重新编辑：重新进入数据连接配置界面，修改配置后重新启动任务。

图 3-21 重试任务



选择重启任务类型后，单击“确认”可重启任务。

数据连接任务详情管理

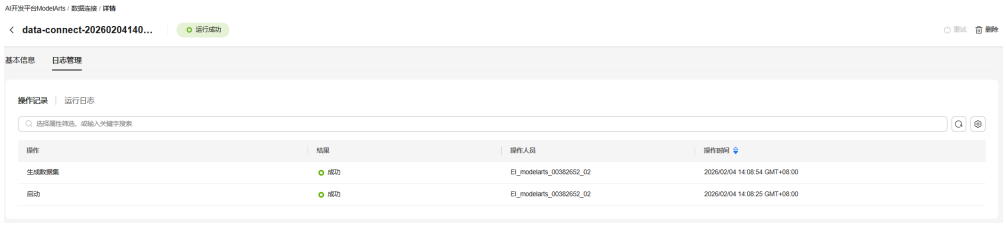
数据任务详情页面展示了当前任务详细信息。在数据连接工作区，单击任意任务名称，就进入该任务的任务详情页面。该页面有基本信息和日志管理两个子页面，以下分别说明两个页面的作用和涉及的操作。

- **基本信息**。该页面列举了该任务的数据集信息、数据配置详情以及生成数据集的链接。在该页面右上角可以删除任务。对于失败的任务，可以单击“重试”按钮重新启动。

图 3-22 基本信息

- **日志管理。**在该界面可以查看“操作记录”和“运行日志”。操作记录会记录当前任务做过的所有操作。运行日志则记录运行过程中的日志。两者都有助于定位在创建任务过程中出现的问题。

图 3-23 日志管理



4 数据精炼

4.1 数据精炼功能说明

功能介绍

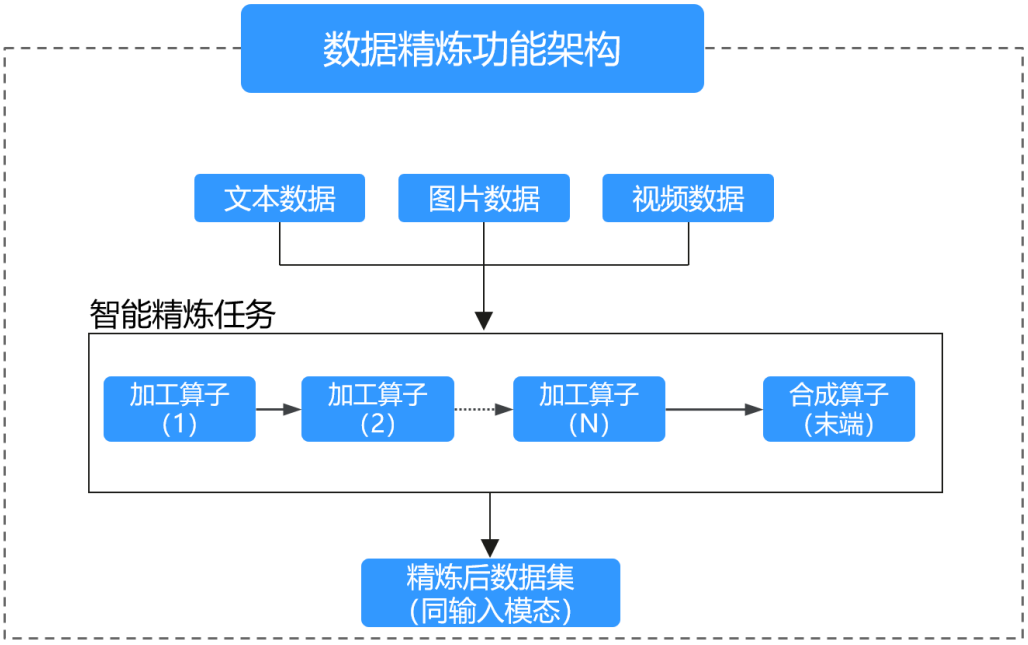
数据精炼是ModelArts数据工程的核心功能模块，旨在解决大模型训练数据准备过程中的“质量”与“数量”双重挑战。它打破了传统数据处理工具的界限，将基于规则的数据加工（清洗、过滤、去重等）与基于大模型的数据合成（改写、扩充、润色等）深度融合。

通过可视化的编数据算子编排，您可以像搭积木一样，将多个加工算子与合成算子串联成一条自动化流水线。系统将按照预设逻辑，对海量原始数据进行层层筛选与优化，最终输出符合训练要求的高质量数据集。

功能架构

数据精炼以文本、图片、视频类数据集作为输入源，构建由多种数据加工算子及合成算子串联编排的智能精炼任务，输出精炼后数据集。具体功能架构参见图4-1。

图 4-1 数据精炼功能架构



核心价值

- **流程统一**：加工与合成一体化编排，无需在多个功能模块间切换，减少中间数据流转，一个任务即可完成从原始脏数据到高质量训练集的全过程。
- **质量提升**：通过多级加工算子层层过滤，确保进入合成环节的数据质量可靠。
- **灵活编排**：支持几十种算子自由组合，满足从简单清洗到复杂增强的各类业务场景。
- **规模扩充**：在清洗后的高质量数据基础上进行合成改写，高效扩充训练数据。
- **效率提升**：可视化算子编排，所见即所得，无需编写处理脚本。
- **可复现性**：使用精炼模板精炼数据，精炼模板可保存、可复用，保证数据处理流程的一致性。

4.2 数据精炼快速入门

业务场景

指定的文本数据集中存在个人敏感信息（电话号码、邮箱信息、车牌号码），现在需要将文本数据的敏感信息做脱敏处理。

通过数据精炼任务，完成敏感信息的脱敏。

前提条件

1. 已注册华为账号并开通华为云，进行了实名认证，且在使用ModelArts前检查账号状态，账号不能处于欠费或冻结状态。
   - **注册华为账号并开通华为云**
   - **进行实名认证**

2. 配置委托访问授权

   ModelArts使用过程中涉及到OBS等服务交互，首次使用ModelArts需要用户配置委托授权，允许访问这些依赖服务。

3. 已申请到数据精炼过程要使用到的计算资源。

约束限制

- 仅西南-贵阳一区域的新版控制台支持。

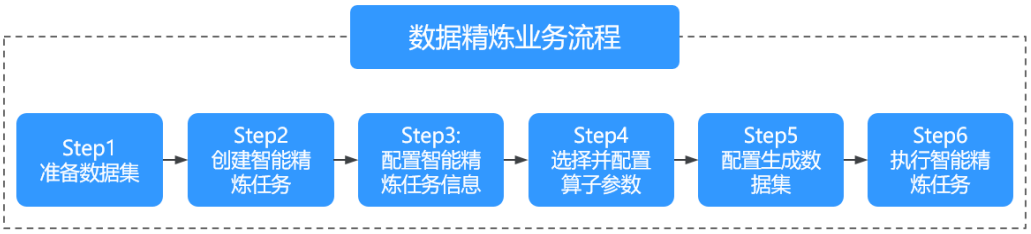
计费说明

数据连接计费涉及到数据存储OBS计费，具体可参考**数据管理计费项**。

创建数据精炼步骤

数据精炼业务流程整体如**图4-2**所示。

图 4-2 数据精炼流程



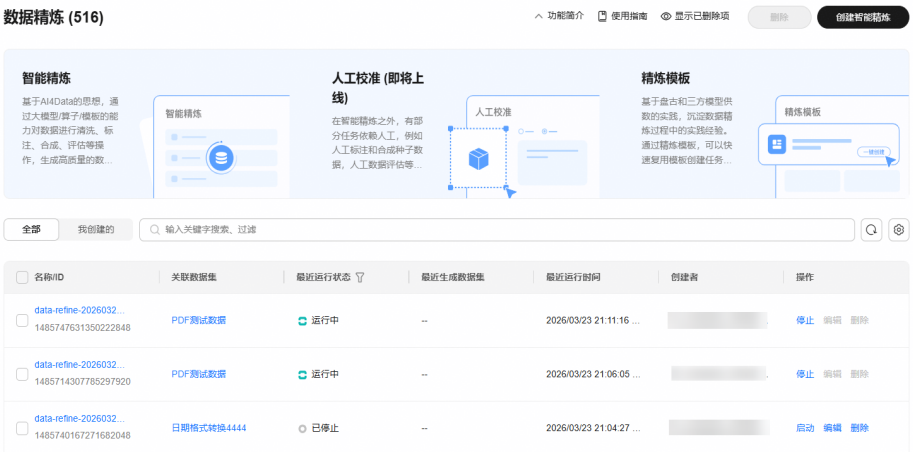
1. 前往**ModelArts管理控制台**。
2. 准备数据集。数据精炼支持**文本、图片、视频**类数据集作为输入数据，数据导入可参考**数据连接**提前将数据导入到ModelArts平台，也可使用ModelArts平台预置数据集。根据本案例场景，可以选择一个包含个人相关信息的文本数据。
3. 在控制台左侧导航栏选择“数据准备 > 数据精炼”，打开“创建智能精炼”工作区，如**图4-3**所示。

   图 4-3 智能精炼

4. 创建智能精炼任务。在**数据精炼**工作区右上方单击“创建智能精炼”按钮，打开“创建智能精炼”配置页配置智能精炼任务相关信息。配置信息如下：
   - 基本信息：配置任务名称和描述信息。任务名称为必选，描述信息为可选。

图 4-4 配置基本信息



说明：

名称：命名默认为data-refine-年月日时分秒，如：data-refine-20260226084902，也可自定义名称，命名要求如下：

- 命名长度：2~64字符。
- 格式要求：以中文、字母**开头**，以中文、字母、数字**结尾**。只允许输入中文、字母、数字、中划线、下划线等字符。

描述：无格式要求，长度不超过200字符，内容可选填。

- 选择数据集。数据集可以选择ModelArts**预置数据**，也可以选择**我的数据**。当前支持输入数据集模态为文本、图片、视频类。本案例选择类型为“单轮问答”的文本类数据集。

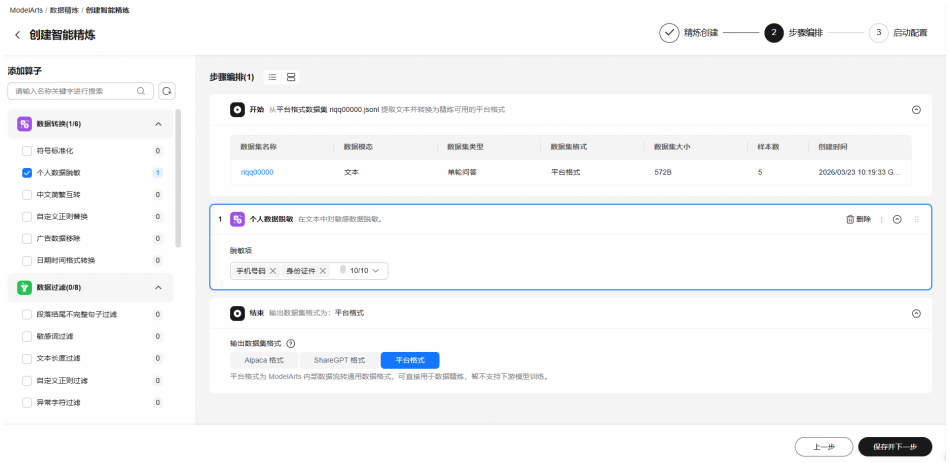
图 4-5 选择数据集

- 选择精炼模板。ModelArts预置了常用数据精炼模板，精炼模板根据使用场景将常用的业务处理算子及各算子的参数配置完毕，您可以直接使用。如果您的使用场景不是精炼模板包含的业务，可不选择精炼模板，直接单击“下一步”。

图 4-6 精炼模板

5. 选择并配置相关算子。本案例中需要对文本做脱敏处理，选择“数据转换 > 个人信息脱敏”算子。选择后，在工作区出现个人数据脱敏的数据算子编排区域。在个人脱敏算子中选择手机号码、邮箱地址、国内车牌号脱敏。在结束节点选择精炼后的数据集格式，支持开源的**Alpaca**、**ShareGPT**格式数据集或平台格式数据集。本案例使用默认的平台格式选择完成后，右下角单击“下一步”。

图 4-7 个人脱敏算子配置

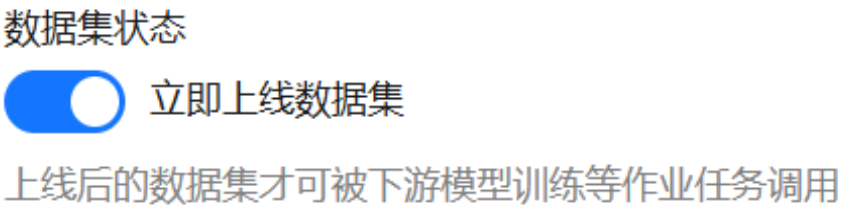


6. 配置生成数据集。需要配置输入数据集名称、存储地址、数据集属性（可选）、描述信息（可选）。

图 4-8 生成数据集

7. 数据集填写完成后，配置“立即上线数据集”。本案例打开数据上线开关。
   - 选择立即上线数据集，生成数据集在“**资产管理 > 数据 >我的数据**”为上线状态，可以被下游模型训练等作业直接使用。
   - 不选择立即上线数据集，生成数据集在“**资产管理 > 数据 >我的数据**”为下线状态，不可被下游模型训练等作业直接使用，需要手动上线数据集后才能使用。

图 4-9 选择“立即上线数据集”



8. 配置“资源配置”。数据算子在处理数据时，需要用到计算资源，针对处理场景和算子类型不同，需要用到CPU或NPU类资源，需要根据选择算子具体确认。本案例中使用公共资源池，并选择CPU资源，会涉及到CPU资源的计费，费用信息请以实际账单为准。配置完毕后单击右下角“确认”后，启动智能精炼任务。

图 4-10 资源配置

9. 智能精炼任务完成后，即可按照要求生成一份将个人数据做过脱敏处理的数据集。生成的数据集可在控制台左侧选择“**资产管理 > 数据 >我的数据**”列表中查看。

4.3 数据精炼使用场景

数据精炼是面向大模型训练的“清洗+合成”一体化数据准备方案。如果您的数据处于“原始粗糙”或“样本匮乏”状态——无论是需要去除HTML标签与乱码的预训练语料清洗，还是需要对少量种子数据进行扩充与润色的SFT指令微调增强，亦或是涉及隐私信息的安全合规脱敏——数据精炼均能通过丰富的数据处理算子编排为一条流水线，将原始数据转化为高质量、高多样性且安全合规的训练数据集。

数据精炼虽然具备强大的数据处理能力和灵活的算子编排方式，但正因为如此，也会让您上手存在一些困难。本文总结了数据精炼的一些常用场景，让您轻松完成数据精炼任务，快速获取高质量的数据，模型开发快人一步。

典型使用场景

根据业务需求不同，数据精炼主要应用于以下几种典型场景，每种场景都配备了推荐的算子组合。您可以根据自己的需求选择不同场景。

- 仅对文本类数据集格式做转换，请选择**场景一：数据集格式转换**。
- 数据质量差，需要清洗，请选择**场景二：原始语料清洗与质量提升**。
- 数据量不足，需要扩充，请选择**场景三：训练数据扩充与增强**。
- 准备SFT微调数据 ，请选择**场景四：指令微调数据准备**。
- 处理图像/视频数据，请选择**场景五：多模态数据统一处理**。
- 数据合规性要求，请选择**场景六：数据合规与安全处理**。

场景一：数据集格式转换

场景描述

ModelArts平台支持多种数据集格式，需要将数据从一种格式转成另外格式的数据集，不做额外的数据处理。

推荐算子编排顺序

开始节点 → 结束节点

预期效果

实现输入数据转为不同格式（Alpaca格式/ShareGPT格式/平台格式）的输出数据。

场景二：原始语料清洗与质量提升

场景描述

企业从互联网爬取、内部系统导出或第三方采购的原始数据，通常存在大量噪声，需要系统性清洗以满足模型训练要求。常见语料问题见表4-1。

表 4-1 典型数据问题

数据问题	表现形式	对模型的影响
数据中存在重复信息。	数据中存在大量相同或相似内容。	导致训练的模型过拟合。
数据中存在乱码噪声。	数据中存在编码错误、异常字符。	污染模型语义理解。
数据中有敏感违规信息。	数据中存在涉政/涉黄/暴力内容。	模型输出合规风险。
数据质量低下。	语句不通、逻辑混乱，句子不完整。	降低模型生成质量。
数据长度不满足要求。	数据过短无意义或过长冗余。	训练效率低下。
数据大杂烩，未分类	各领域的数据杂糅，没有按领域分类。	需要专门对领域数据分类，影响训练效率。

推荐算子编排顺序

原始语料 → [符号标准化] → [去重算子] → [敏感词过滤] → [文本长度过滤] → [段落结尾不完整句子移除] → [色情文本检测] → [涉政文本检测] → [辱骂文本检测算子] → [预训练文本分类] → 清洗后数据

预期效果

- 数据重复率降低90%以上。
- 低质量样本有效去除。
- 敏感违规内容100%过滤。
- 输出数据可直接用于训练或进入下一步合成环节。
- 输出数据能够按照不同领域分类。

场景三：训练数据扩充与增强

高质量标注数据获取成本高，现有数据量不足以训练出效果良好的模型。

适用情况

- 垂直领域数据稀缺。
- 标注成本过高。
- 需要快速扩充数据规模。
- 数据多样性不足。

推荐算子编排顺序

原始数据 → 数据清洗 → 数据生成 → 扩充后数据

表 4-2 使用算子

算子	作用	配置建议
数据清洗	确保种子数据质量	严格筛选标准
数据生成	生成多样化表达	选择合适的改写策略，生成多样化数据。

预期效果

- 数据规模扩充3-10倍。
- 保持语义一致性。
- 提升表达多样性。

注意事项

- 合成算子需放在工作流末端。
- 仅支持同模态数据合成。

场景四：指令微调数据准备

准备用于大模型指令微调（SFT）的高质量数据集。

适用情况

- 通用助手模型微调。
- 垂直领域模型定制。
- 对话能力优化。
- 任务型模型训练。

推荐算子编排流程

原始指令数据 → 数据清洗 → 文本生成（可选） → 生成数据集

表 4-3 数据格式处理

输入格式	处理方式	输出格式
非结构化文本	格式转换算子	Alpaca/ShareGPT
已有Alpaca	质量筛选+改写	优化后Alpaca
已有ShareGPT	质量筛选+改写	优化后ShareGPT

质量控制要点

- 指令明确性检查
- 回答准确性验证
- 格式一致性确保

场景五：多模态数据统一处理

处理包含图像、视频等多种模态的数据集。

适用情况

- 视频理解数据整理

推荐算子编排流程（以图像为例）

```
图像数据集 → 图片去重 → 图片提取 → 图片元数据过滤 → 图像检测 → 处理后数据
```

表 4-4 各模态处理要点

模态	关键处理	注意事项
图像	尺寸、格式、质量	分辨率统一
视频	帧率、分辨率、片段	视频编码统一

重要约束

每种模态需单独创建处理任务。

场景六：数据合规与安全处理

确保训练数据符合法规要求和企业安全策略。

适用情况：

- 个人信息保护（GDPR/个保法）。
- 敏感内容过滤。

推荐算子编排流程

```
原始数据 → 敏感词过滤 → 合规数据
```

使用算子：

算子	作用	合规要求
敏感词过滤	过滤个人信息敏感内容。	符合个人隐私要求。

4.4 创建数据精炼

使用场景

数据精炼是面向大模型训练的“清洗+合成”一体化数据准备方案。如果您的数据处于“原始粗糙”或“样本匮乏”状态——无论是需要去除HTML标签与乱码的预训练语料清洗，还是需要对少量种子数据进行扩充与润色的SFT指令微调增强，亦或是涉及隐私信息的安全合规脱敏——数据精炼均能通过丰富的数据处理算子编排为一条流水线，将原始数据转化为高质量、高多样性且安全合规的训练数据集。

前提条件

1. 已注册华为账号并开通华为云，进行了实名认证，且在使用ModelArts前检查账号状态，账号不能处于欠费或冻结状态。
   - **注册华为账号并开通华为云**
   - **进行实名认证**
2. 配置委托访问授权

   ModelArts使用过程中涉及到OBS等服务交互，首次使用ModelArts需要用户配置委托授权，允许访问这些依赖服务。
3. 已申请到数据精炼过程要使用到的计算资源。

计费说明

数据连接计费涉及到数据存储OBS计费，具体可参考**数据管理计费项**。

约束限制

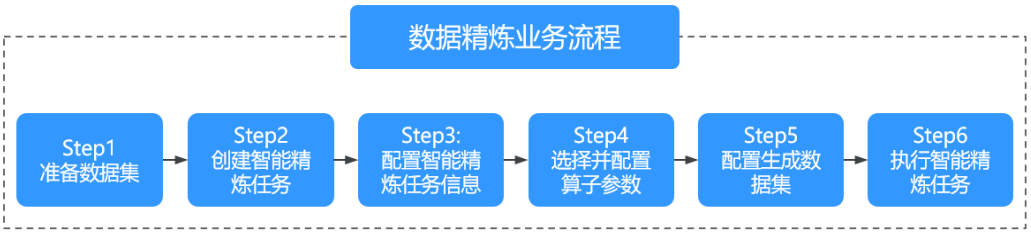
- 仅西南-贵阳一区域的新版控制台支持。
- **合成算子位置限制**：合成算子**必须且只能**放置在工作流的**最后一个节点**。不支持插入到加工算子中间，也不支持后接其他过滤算子。
- **模态限制**：
  - 仅支持**同模态**合成（如文本输入 -> 文本输出）。
  - **不支持跨模态**生成（如输入文本生成问答对、输出文本生成图像等）。
- **功能限制**：
  - 不支持自定义合成指令（Prompt），不支持模板功能。
  - 不支持合成任务的在线调测。
  - 合成算子输出字段固定，但会自动保留输入数据集中的原始字段。
- **数据量与质检**：
  - 不支持用户自定义合成数据的输出条数（系统根据输入自动处理）。
  - 不支持对合成结果进行自动质检过滤。

- 合成结果将输出到新数据集，不支持自动与原始数据集合并。

创建数据精炼步骤

数据精炼业务流程整体如图4-2所示。

图 4-11 数据精炼流程

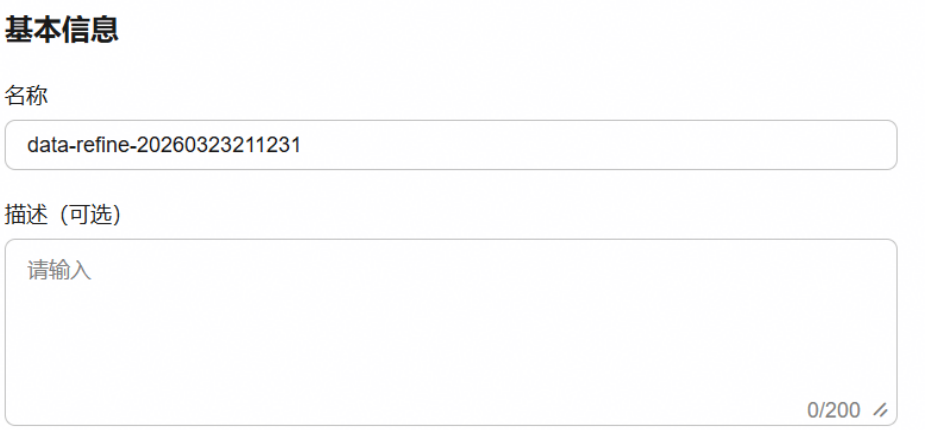


1. 前往ModelArts管理控制台。
2. 准备数据集。数据精炼支持**文本、图片、视频**类数据集作为输入数据，具体数据导入可参考**数据连接**提前将数据导入到ModelArts平台，也可使用ModelArts平台预置数据集。
3. 在控制台左侧导航栏选择“数据准备 > 数据精炼”，打开“创建智能精炼”工作区，如图4-3所示。

图 4-12 智能精炼

4. 创建智能精炼任务。在**数据精炼**工作区右上方单击“创建任务”按钮，打开“创建智能精炼”配置页配置智能精炼任务相关信息。配置信息如下：
   - 基本信息：配置任务名称和描述信息。任务名称为必选，描述信息为可选。

图 4-13 配置基本信息



说明：

名称：命名默认为data-refine-年月日时分秒，如：data-refine-20260226084902，也可自定义名称，命名要求如下：

- 命名长度：2~64字符。
- 格式要求：以中文、字母**开头**，以中文、字母、数字**结尾**。只允许输入中文、字母、数字、中划线、下划线等字符。

描述：无格式要求，长度不超过200字符，内容可选填。

- 选择数据集。数据集可以选择ModelArts平台**预置数据**，也可以选择**我的数据**。当前支持输入数据集模态为**文本、图片、视频**类。

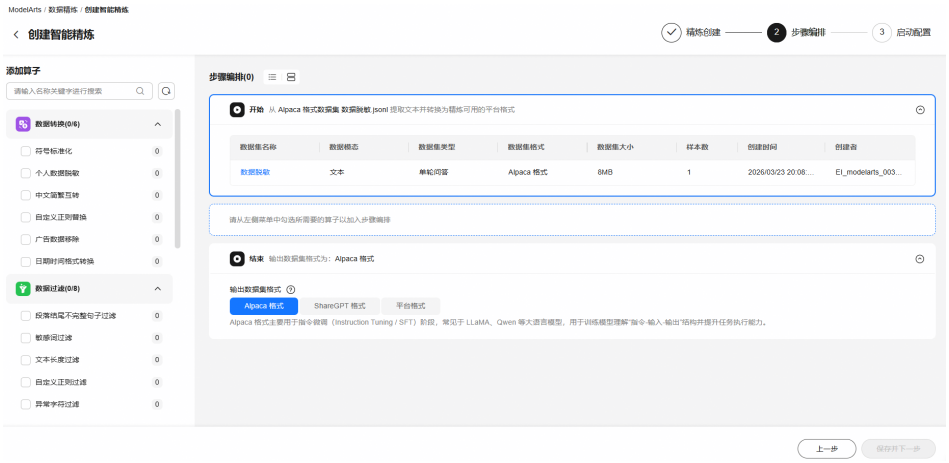
图 4-14 选择数据集

- 选择精炼模板。ModelArts预置了常用数据精炼模板，精炼模板根据使用场景将常用的业务处理算子及各算子的参数配置好，您可以直接使用，如需进一步了解精炼模板，请参考**管理数据精炼模板**。如果您的使用场景不是精炼模板包含的业务，可不选择精炼模板，直接单击“下一步”。

图 4-15 精炼模板

5. 选择并编排数据算子。需要根据场景选择不同的算子做编排，如果对算子使用场景有疑问，请参考**数据精炼使用场景**章节选择，如果需要了解算子详细用法，请参考**管理数据精炼算子**。选择算子后，在工作区出现算子编排区域，配置算子顺序及算子参数后。右下角单击“保存并下一步”。

图 4-16 算子编排



重要：数据编排是数据精炼最重要也最为复杂的处理过程，有很多需要注意的场景和约束。特别是针对部分文本类（**单轮对话、单轮对话带人设、多轮对话、多轮对话带人设**）数据，**开始节点**和**结束节点**不仅负责处理数据的输入输出，还兼具数据格式转换功能。以下针对有数据格式转换的场景做场景说明：

- 任意格式数据集进入开始节点后，均需要转换为平台格式数据集，供后续算子处理。待各种类型算子处理完毕后，结束算子将输出数据集默认会转化为输入数据集同格式数据集。
- 结束节点可以配置输出数据集格式为任意格式数据集，供用户自己选择。
- 如果输入节点和输出节点之间未添加其余任何算子，需按照如下两种情况处理：
  - 结束节点设置输出数据集格式与开始节点输入数据集格式一致。此时相当于对数据集未做任何操作，下一步按钮置灰，无法进行下一步配置。
  - 结束节点设置输出数据集格式与开始节点输入数据集格式不同。此时相当于仅对数据集做格式转换，可进行下一步配置，完成数据精炼任务后续配置。
- 单击“保存并下一步”后，当前数据精炼编排任务包含编排步骤及之前的配置都会被保存下来。如果后续任务未完成，可在下次打开该数据精炼任务后，继续完成后续配置。

6. 配置生成数据集。需要配置输入数据集名称、存储地址、数据集属性（可选）、描述信息（可选）。

图 4-17 生成数据集

说明：

数据集名称：

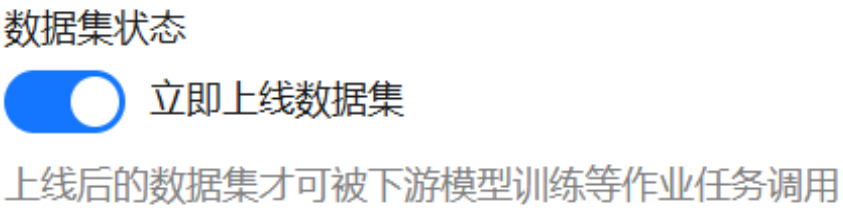
- 命名长度：2~63字符。
- 格式要求：以中文、字母**开头**，以中文、字母、数字**结尾**。只允许输入中文、字母、数字、中划线、下划线字符。

数据集属性：可选字段，支持配置标签。可以按照行业、语言维度配置标签，也可自定义标签。

描述：可选字段，无格式要求，长度不超过200字。

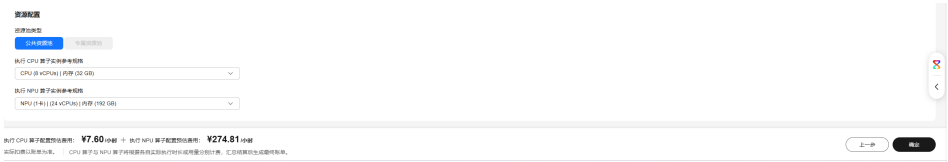
7. 数据集填写完成后，配置“立即上线数据集”。
   - 选择**立即上线数据集**，生成数据集在“**资产管理 > 数据 >我的数据**”为上线状态，可以被下游模型训练等作业直接使用。
   - 不选择**立即上线数据集**，生成数据集在“**资产管理 > 数据 >我的数据**”为下线状态，不可被下游模型训练等作业直接使用，需要手动上线数据集后才能使用。

图 4-18 选择“立即上线数据集”



8. 配置“资源配置”。数据算子在处理数据时，需要用到计算资源，针对处理场景和算子类型不同，需要用到CPU或NPU类资源，需要根据选择算子具体确认。使用公共资源池，并选择CPU资源或NPU资源，会涉及到CPU或NPU资源的计费，费用信息请以实际账单为准。配置完毕后单击右下角“确认”后，启动智能精炼任务。

图 4-19 资源配置



9. 智能精炼任务完成后，生成的数据集可在控制台左侧选择“**资产管理 > 数据 >我的数据**”列表中查看。

最佳实践：算子编排设计原则

原则1：先清洗，后处理。

建议顺序：去重 → 格式化 → 过滤 → 增强 → 合成

原则2：减少数据在前，扩充数据在后。先用过滤算子减少数据量，再用合成算子扩充，可提升整体处理效率。

原则3：合成放末端，合成算子只能作为最后一个处理步骤。

原则4：保持模态一致，整个工作流处理同一类型数据，不跨模态。

推荐算子编排模板

模板一：基础数据清洗

输入 → 格式校验 → 去重 → 长度过滤 → 输出

模板二：数据清洗+质量提升

输入 → 格式校验 → 去重 → 敏感词过滤 → 质量评分筛选 → 输出

模板三：数据清洗+合成扩充

输入 → 去重 → 敏感词过滤 → 质量筛选 → 问答改写合成 → 输出

模板四：全流程精炼

输入 → 格式转换 → 去重 → 敏感词过滤 → 质量评分 → 长度筛选 → 改写合成 → 输出

4.5 管理数据精炼

4.5.1 管理数据精炼任务

约束限制

- 仅西南-贵阳一区域的新版控制台支持。

功能说明

在数据精炼任务工作区列举了所有的数据精炼任务列表，可以查看数据精炼任务的**名称/ID、关联数据集、最近运行状态、最近生成数据集、最近运行时间、创建者**以及支持的一些操作。数据精炼标题可以看到当前精炼任务已选择任务**数量/总数量**。如**图4-20**所示。本文将详细介绍如何管理数据精炼任务。

图 4-20 查看数据精炼工作区

展示/隐藏功能简介

打开数据精炼工作区，最上方默认会展示数据精炼的功能介绍，用户能快速感知数据精炼的功能亮点及业务范围，快速上手无难度。

图 4-21 打开功能介绍

对于经验丰富的用户，也可单击功能介绍，关闭该功能。

图 4-22 关闭功能介绍

数据精炼任务管理

数据精炼任务支持**过滤、搜索、启动、停止、重试、编辑、删除**等操作。以下分别讲解如何操作。

- **数据精炼任务过滤和搜索。**在数据精炼任务很多的情况下，支持按照**ID、最近运行状态、创建者**等维度过滤想要的任务，便于快速找到目标任务。也可单击“我创建的”按钮，只列举当前登录用户创建的数据连接任务。

图 4-23 数据集过滤

- **数据精炼任务启动。**

  对**最近运行状态**为**数据集生成成功、未启动、已停止**的数据精炼任务，如果想再次运行该任务，可单击操作列“启动”按钮或任务详情页右上角“启动”按钮，在页面右侧弹出的对话框，重新修改精炼配置，启动该精炼任务。如**图4-24**所示。具体可参考**创建数据精炼步骤**的步骤6及之后步骤完成配置。

图 4-24 启动智能精炼

  配置完成后，单击“确定”，启动智能精炼任务。

- **数据精炼任务停止。**

创建好的数据精炼任务，如果任务还在运行中，可单击操作列“停止”按钮或任务详情页右上角“停止”按钮，在弹出的对话框中“确认”后，该任务停止运行。

图 4-25 停止数据精炼任务

- **数据精炼任务重试。**

  对于运行失败的任务，支持“重试”操作重新运行该任务。可以在“操作”列单击“重试”按钮或任务详情页面右上角单击“重试”后，在页面右侧弹出的对话框，重新修改精炼配置，重试该精炼任务。如图4-26所示。具体可参考**创建数据精炼步骤**的步骤6及之后步骤完成配置。

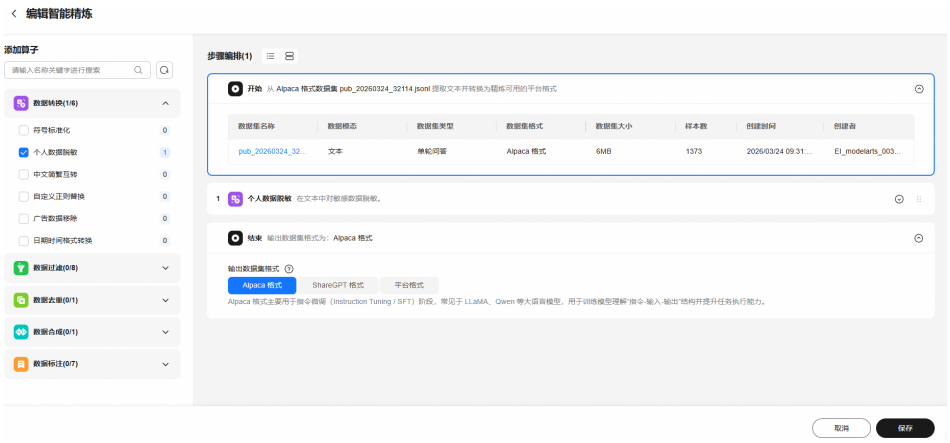
图 4-26 启动智能精炼

配置完成后，单击“确定”，启动智能精炼任务。

- **数据精炼任务编辑。**

  最近运行状态不是“运行中”状态的数据精炼任务，均可单击编辑，重新编辑该数据任务的数据算子参数或调整顺序。单击“编辑”后，弹出编辑智能精炼配置页，调整算子顺序或参数后，单击“保存”完成任务的编辑。也可单击“取消”按钮，取消编辑。如图8 编辑智能精炼。

图 4-27 编辑智能精炼



- **数据精炼任务删除。**

  对于列表中创建的数据精炼任务，支持“删除”操作。可以在“操作”列单击“删除”按钮，在弹出删除对话框选择“确定”后，该任务将被删除。删除后的任务不是彻底删除，为避免误删，如果还想再继续使用，可以恢复任务。已删除的任务在工作区右上角单击“显示已删除项”（如图5 显示已删除任务）后，在任务清单有可以看到已删除的任务名称后有“已删除”标签，如图3-18所示。对于已删除的任务，可以选择“彻底删除”，如图3-19所示。彻底删除后的任务不可恢复。

图 4-28 显示已删除任务

图 4-29 删除任务

图 4-30 已删除任务的“已删除”标签

图 4-31 任务“彻底删除”

- **数据精炼任务批量删除**

  勾选数据精炼任务名称前的复选框，选择要删除的任务后，单击右上角“删除”按钮，可以批量删除数据。

图 4-32 批量删除

数据精炼任务详情管理

数据精炼任务详情页面展示了当前任务详细信息。在数据精炼工作区，单击任意任务名称，就进入该精炼任务的任务详情页面。在该页面右上角可以根据任务的不同状态选择**启动、重试、删除、停止**等按钮，单击后操作参见**数据精炼任务管理**对应操作。

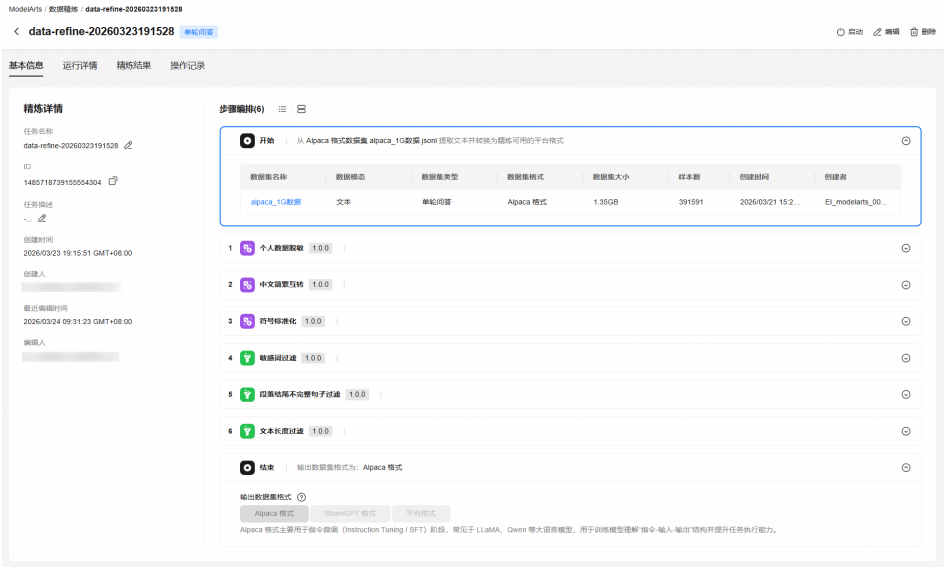
数据任务详情有**基本信息、运行详情、精炼结果、操作记录**四个子页面。以下分别说明页面的作用和涉及的操作。

基本信息

基本信息左侧列举了**精炼详情**，包含**任务名称、ID、任务描述、创建时间、创建人、最近编辑时间、编辑人等信息。任务名称**和**任务描述**支持修改。

基本信息右侧列举了该任务的使用的数据算子的编排详情。

图 4-33 基本信息



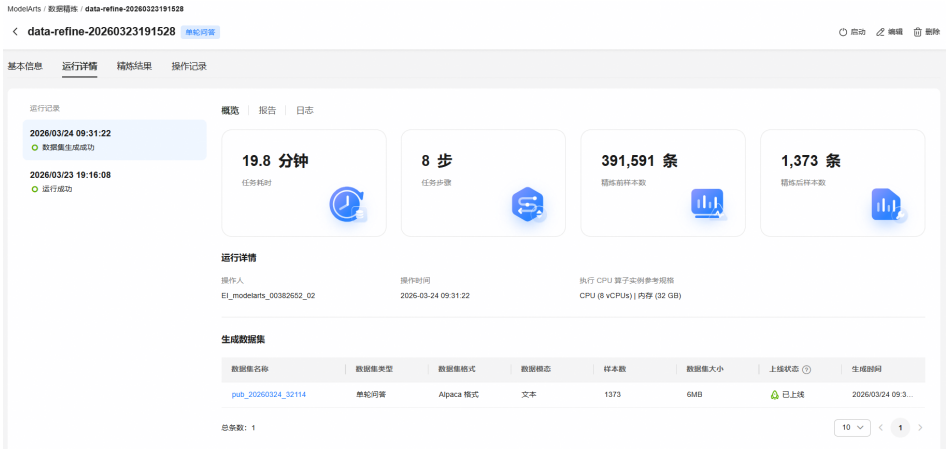
运行详情

运行详情左侧列举了该任务运行记录，包含运行时间及运行结果状态。

运行详情右侧列举运行任务的概览、报告、日志。分别说明如下：

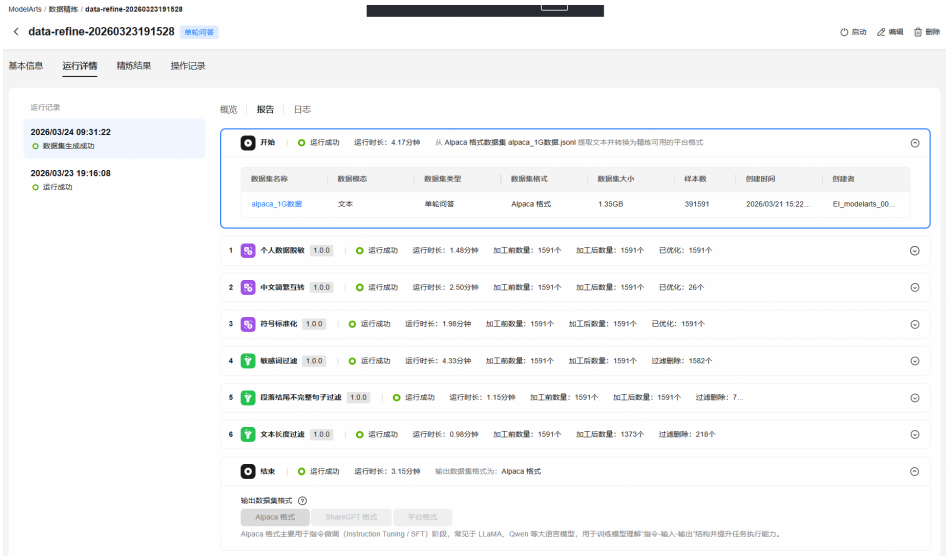
- **概览**。展示每次运行记录的任务耗时、任务步骤、精炼前样本数量、精炼后样本数量。运行详情包含了操作人、操作时间、任务使用的资源的详情。生成数据集包含了生成数据集的链接及其余信息，如图4-34所示。

  图 4-34 概览

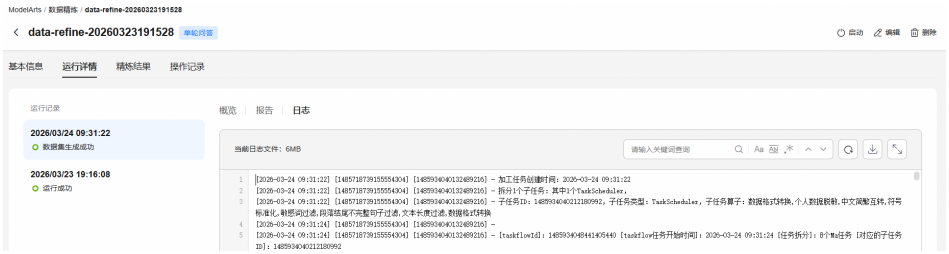
- **报告**。展示每次运行记录的编排步骤各数据算子运行状态，算子运行时长、算子处理前后数据的样本数量及算子优化命中数量。如图4-35所示。通过报告可以直观定位各个算子处理的数据是否符合预期。

图 4-35 报告



- **日志**。展示每次运行记录的数据精炼任务运行过程中的日志记录，通过日志能够快速定位数据精炼任务出现的问题。日志界面支持按照正则匹配关键字，查找关键日志。如图4-36所示。

图 4-36 日志



精炼结果

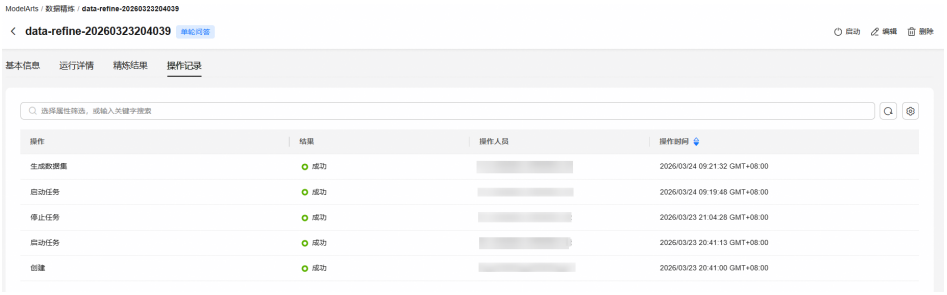
展示通过数据精炼任务生成的数据集信息，包含数据集在数据资产中的链接。具体如图4-37所示。

图 4-37 精炼结果

操作记录

记录数据精炼任务所有的操作记录，便于查看当前任务的操作状态。具体如图4-38所示。

图 4-38 操作记录



4.5.2 管理数据精炼模板

功能简介

数据精炼模板是ModelArts平台沉淀的数据处理最佳实践集合。它将复杂的算子组合按特定业务逻辑预先编排，封装为一键可用的标准化工作流。

它的核心作用在于：

1. 无需从零开始研究算子顺序与参数，直接复用资深算法工程师验证过的成熟方案，确保数据处理流程的专业性与合理性。
2. 用户只需在创建任务时选择匹配业务场景的模板，系统即可自动加载完整的算子链路。您既可以直接运行，也能在此基础上根据数据特性进行微调，将原本小时级的数据准备工作缩短至分钟级。
3. 通过模板化机制，规范团队内部的数据处理标准，避免因个人配置差异导致的数据质量波动，确保输出的训练集始终维持在高水平。

预置模板说明

ModelArts平台当前提供以下预置模板，覆盖文本、图片、视频相关数据的处理流程，具体使用流程参见表4-5。

表 4-5 预置模板清单

模板名称	支持数据集模态	支持数据集类型	模板使用场景	涉及算子
WORD处理流程	文本	文档	使用预置算子，对word文件进行提取和处理，生成预训练文本数据。	WORD内容提取
				个人数据脱敏
				中文简繁互转
				符号标准化
				敏感词过滤
				段落结尾不完整句子过滤
				N-gram特征过滤

模板名称	支持数据集模态	支持数据集类型	模板使用场景	涉及算子
				文本长度过滤
				预训练文本分类
PDF处理流程	文本	文档	使用预置算子，对PDF文件进行提取和处理，生成预训练文本数据。	PDF内容提取
				个人数据脱敏
				中文简繁互转
				符号标准化
				敏感词过滤
				段落结尾不完整句子过滤
				N-gram特征过滤
				文本长度过滤
				预训练文本分类
预训练文本处理流程	文本	预训练文本	使用预置算子，对预训练文本进行处理，生成清洗后的预训练文本数据。	中文简繁互转
				符号标准化
				N-gram特征过滤
				文本长度过滤
				预训练文本分类
图片处理流程	图片	图片	使用预置算子，对图片进行提取、去重、打标和过滤，以及生成摘要，生成处理后的图文数据。	图文提取
				图片元数据过滤
				图片去重
				危情图像检测
				色情图像检测
				暴恐图像检测
视频处理流程	视频	视频	使用预置算子，对视频进行镜头拆分、去重、打标和过滤，生成处理后的视频数据。	镜头拆分
				视频元数据过滤
				视频宽高比过滤
				色情视频检测
				暴恐视频检测
				视频涉政检测

模板名称	支持数据集模态	支持数据集类型	模板使用场景	涉及算子
				运动幅度评分
				美学评分
单轮问答处理流程	文本	单轮问答	使用预置算子，对单轮问答数据进行处理，生成清洗后的单轮问答数据。	个人数据脱敏
				中文简繁互转
				符号标准化
				敏感词过滤
				段落结尾不完整句子过滤
				N-gram特征过滤
				文本长度过滤
问答排序处理流程	文本	问答排序	使用预置算子，对问答排序数据进行处理，生成清洗后的问答排序数据。	个人数据脱敏
				中文简繁互转
				符号标准化
				敏感词过滤
				段落结尾不完整句子过滤
				文本长度过滤
多轮问答处理流程	文本	多轮问答	使用预置算子，对多轮问答数据进行处理，生成清洗后的多轮问答数据。	个人数据脱敏
				中文简繁互转
				符号标准化
				敏感词过滤
				段落结尾不完整句子过滤
				文本长度过滤
偏好优化处理流程	文本	偏好优化	使用预置算子，对偏好优化数据进行处理，生成清洗后的偏好优化数据。	个人数据脱敏
				中文简繁互转
				符号标准化
				敏感词过滤
				段落结尾不完整句子移除

模板名称	支持数据集模态	支持数据集类型	模板使用场景	涉及算子
				文本长度过滤

4.6 管理数据精炼算子

4.6.1 预置数据精炼算子

数据精炼算子分为**加工算子**和**合成算子**两大类，通过算子的组合编排实现完整的数据处理流程。

表 4-6 数据精炼算子清单

算子类型	算子分类	算子名称	算子描述
文本类加工算子	无分类	开始节点	在数据精炼编排步骤中作为首节点接收待精炼数据集。针对部分文本类（**单轮对话、单轮对话带人设、多轮对话、多轮对话带人设**）数据，具备数据转换的功能。转换规则如下： • 平台格式数据集无需转换，直接进入精炼流程。 • 非平台格式数据（Alpaca格式/ShareGPT格式）进入开始节点，均需要转化为平台格式后，再做后续处理。
		结束节点	在数据精炼编排步骤中作为结束节点输出待精炼后的数据集。针对**部分文本类（单轮对话、单轮对话带人设、多轮对话、多轮对话带人设）**数据，具备数据转换的功能。转换规则如下： • 开始节点输入的数据集为平台格式，结束节点默认输出平台格式数据集。 • 开始节点输入的数据集为非平台格式（Alpaca格式/ShareGPT格式）数据，结束节点默认输出对应非平台格式数据。 • 结束节点可自行选择和开始节点不同格式数据集作为最终输出的数据集格式。
	数据提取	WORD内容提取	从Word文档中提取文字，并保留原文档的目录、标题和正文等结构，不保留图片、表格、公式、页眉、页脚。

算子类型	算子分类	算子名称	算子描述
		CSV内容提取	从CSV文件中读取所有文本内容，并按该文件内容类型模板KEY值生成匹配的JSON格式数据。
		PDF内容提取	从PDF中提取文本，转化为结构化数据，支持文本、表格、公式等内容提取。
	数据转换	个人数据脱敏	对文本中的手机号码、身份证件、邮箱地址、URL链接、国内车牌号、IP地址、MAC地址、IMEI、护照、车架号等个人敏感信息进行数据脱敏，或直接删除敏感信息。
		中文简繁互转	将中文简体和中文繁体进行转换。
		符号标准化	查找文本中携带的非标准化符号进行标准化、统一化转换。 ● 统一空格：将所有Unicode空格（如U+00A0、U+200A）转换为标准空格（U+0020）。 ● 全角转半角：将文本中的全角字符转换为半角字符。 ● 标点符号归一化，支持统一格式的符号如下： – {"? ": "\?\? "} – {"[":"〖"} – {"]":"〗"} ● 数字符号归一化，例如将⓪□□□□统一为0.。支持统一格式的符号如下： – {"0.": "⓪□□□□"} – {"1.": "①|⑴|㈠|1.|⓵|❶|❶|①"} – {"2.": "②|⑵|㈡|2.|⓶|❷|❷|②"} – {"2.": "②|⑵|㈡|2.|⓶|❷|❷|②"} – {"3.": "③|⑶|㈢|3.|⓷|❸|❸|③"} – {"4.": "④|⑷|㈣|4.|⓸|❹|❹|④"} – {"5.": "⑤|⑸|㈤|5.|⓹|❺|❺|⑤"} – {"6.": "⑥|⑹|㈥|6.|⓺|❻|❻|⑥"} – {"7.": "⑦|⑺|㈦|7.|⓻|❼|❼|⑦"} – {"8.": "⑧|⑻|㈧|8.|⓼|❽|❽|⑧"} – {"9.": "⑨|⑼|㈨|9.|⓽|❾|❾|⑨"} – {"10.": "⑩|⑽|㈩|10.|⓾|❿|❿|⑩"}

算子类型	算子分类	算子名称	算子描述
		自定义正则替换	数据条目不变下，使用自定义正则表达式替换文本内容。 示例如下： • 去除“参考文献”以及之后的内容：**\n参考文献[\s\S]*** • 针对pdf的内容，去除“0 引言”之前的内容，引言之前的内容与知识无关：**[\s\S]{0，10000}0 引言** • 针对pdf的内容，去除“1.1Java简介”之前的与知识无关的内容：**[\s\S]{0，10000} 1\. 1Java简介**
		日期时间格式转换	自动识别日期、时间、星期，同时根据选择的格式进行统一转换。
		广告数据移除	按照句子的过滤粒度，删除文本中包含广告数据的句子。
	数据过滤	异常字符过滤	查找数据集每一条数据中携带的异常字符，并将异常字符替换为空值，数据条目不变。 • 不可见字符，例如U+0000-U+001F。 • 表情符□□。 • 网页标签符号<style></style>。 • 特殊符号，例如● ■ ◆。 • 乱码和无意义的字符□□□□□。 • 特殊空格：[\u2000-\u2009]
		自定义正则过滤	删除或保留符合自定义正则表达式的数据。
		自定义关键词过滤	剔除包含关键词的数据。
		段落结尾不完整句子过滤	按照句子的过滤粒度，自动识别段落结尾处的内容是否完整，如果不完整，则删除。
		敏感词过滤	对文本中涉及黄色、暴力、政治等敏感数据进行自动检测和过滤。
		文本长度过滤	按照设置的文本长度，保留长度范围内的数据。

算子类型	算子分类	算子名称	算子描述
		N-gram特征过滤	用于判断文档重复度，根据特征N值计算文档内词语按N值组合后的重复此时，可通过以下两种算法比较结果是否大于特征阈值，大于特征阈值的文档删除。 • top-gram过滤：计算**重复最多**的gram占总长度的比例，大于特征阈值则删除。 • gram重复率过滤：计算**所有重复**的gram占总长度的比例，大于特征阈值则删除。
		句子特征过滤	该算子将文档中的标点符号作为句子分隔符，统计每句字符长度，如果文档平均字符长度大于设置字符长度，则保留，反之则删除整篇文档。根据如下特征过滤： • 待保留的平均句长。
	数据标注	违禁文本检测	违禁内容检测算子通过对输入中文文本内容分析，最终返回文本中是否含有违禁内容的JSON结构化结果。
		个人隐私识别	个人隐私内容检测算子通过对输入中文文本内容分析，最终返回文本中是否含有个人隐私内容的JSON结构化结果。
		垃圾内容文本检测	垃圾内容检测算子通过对输入中文文本内容分析，最终返回文本中是否含有垃圾内容的JSON结构化结果。
		广告文本检测	垃圾广告内容检测算子通过对输入中文文本内容分析，最终返回文本中是否含有垃圾广告内容的JSON结构化结果。
		色情文本检测	色情内容检测算子通过对输入中文文本内容分析，最终返回文本中是否含有色情内容的JSON结构化结果。
		辱骂文本检测	辱骂内容检测算子通过对输入中文文本内容分析，最终返回文本中是否含有辱骂内容的JSON结构化结果。
		涉政文本检测	政治敏感内容检测算子通过对输入中文文本内容分析，最终返回文本中是否含有政治敏感内容的JSON结构化结果。
		预训练文本分类	针对预训练文本进行内容分类，例如新闻、教育、健康等类别，支持分析语种包括：中文、英文。

算子类型	算子分类	算子名称	算子描述
文本合成类算子	数据合成	数据生成	支持从单一样本生成相似问答、为问答注入特定人设角色，并可一键调整问答难度，实现数据的规模化定制合成。
视频类加工算子	数据提取	视频时长切分	将源视频切分成固定时长的小视频，固定时长可配置，范围为1-5分钟。
		镜头拆分	根据视频中的镜头场景变化将长视频拆分为短视频片段，如果某个镜头片段的长度超过设定的时间阈值，该镜头片段将按时长进行进一步拆分。
	数据转换	视频裁剪	视频裁剪是裁剪掉视频中不必要的元素，例如字幕、Logo、水印、边框和密集文字，同时过滤掉那些裁剪后面积比例超出预设阈值的视频文件；使用前需要先执行字幕、logo、水印、边框、密集文字识别算子。
	数据过滤	视频元数据过滤	根据视频元数据（帧率、分辨率和视频时长）进行过滤，仅保留符合选定条件的视频。注：电影标准帧率为24或30FPS。
		视频宽高比过滤	根据视频的宽高比进行过滤。宽高比是指视频图像的宽度和高度之间的比率。
	数据标注	色情视频检测	给色情视频内容打标签
		暴恐视频检测	给暴恐视频内容打标签
		视频涉政检测	给涉政视频内容打标签
		运动幅度评分	通过计算每个像素在每一帧中的移动范围进行评分，识别运动幅度过快（如＞100光流）或过慢（如≤2光流）的视频，数值越大表示运动越快。
		美学评分	从内容（吸引人，清晰度）、构图（目标物位置良好）、颜色（有活力，令人愉悦）、光线（光线明显有对比度）、轨迹（连续、稳定）等维度评价视频美感得分。分值范围(0, 1)，数值越高美感越好，评分>0.95可视为视频基础质量较高的视频。
		水印识别	识别视频中是否包含水印。
		字幕识别	识别视频中是否包含字幕。

算子类型	算子分类	算子名称	算子描述
		视频黑边识别	识别视频中是否包含黑边。
		密集文字识别	识别视频中是否包含密集文字，超出密集文字面积占比阈值的视频可视为密集文字视频，一般默认裁剪面积占比≥7%为密集文字视频。
		视频分类	通过算子返回视频的标签分类，L1存在10类，L2级别检测39类，L3级别检测93类，L4存在2219类。
		视频摘要生成（简略）	通过对视频进行抽帧，通过模型推理生成简短的视频摘要描述。
		视频摘要生成（详细）	通过对视频进行抽帧，通过模型推理生成详细的视频英文摘要描述。
		视频中文摘要生成（详细）	通过对视频进行抽帧，通过模型推理生成详细的视频中文摘要描述。
		姿态检测	通过对视频抽8帧，模型分别对图片进行标记关键点，输出任务bbox框和关键点坐标，通过对坐标的计算判断视频中是否存在人物。
		镜头运动描述	模型通过对视频进行抽帧进行光流计算与推理，输出视频的镜头类型。
图片类加工算子	数据提取	图文提取	提取图文压缩包中的JSON文本和图片，并对图片进行结构化解析（BASE64编码），方便图文加工算子使用。
	数据过滤	图片元数据过滤	基于图片宽、高、文件大小、宽高比阈值进行图片/图文数据清洗。
		图片去重	通过把图片结构化处理后，过滤重复的图片/图文对数据。
	数据标注	色情图像检测	给图像算子打标签。
		危情图像检测	给危情图片内容打标签
		暴恐图像检测	过滤暴恐图像。

开始节点

- 适用的文件格式：适用于所有类型数据集，但针对“文本类 > 单轮对话、单轮对话带人设、多轮对话、多轮对话带人设”数据具备数据格式转换能力。

- 说明：所有格式数据集，在开始节点处理后都会转化为平台格式数据集。

  针对部分文本类（**单轮对话、单轮对话带人设、多轮对话、多轮对话带人设**）数据，具备数据格式转换的功能。转换规则如下：

  - 平台格式数据集无需转换，直接进入精炼流程。
  - 非平台格式数据（Alpaca格式/ShareGPT格式）进入开始节点，均需要转化为平台格式，供后续数据算子处理。
- 参数配置样例：

  无。
- 转换样例：

  输入节点处理前数据集格式：平台格式/Alpaca格式/ShareGPT格式。

  输入节点处理后数据集格式：平台格式。

结束节点

- 适用的文件格式：适用于所有类型数据集，但针对“**文本类 > 单轮对话、单轮对话带人设、多轮对话、多轮对话带人设**”数据具备数据格式转换能力。您也可以自主选择输出不同格式。
- 说明：

  数据集完成数据精炼后，结束节点能够对指定数据类型数据做数据格式转换。转换规则如下：

  - 开始节点输入的数据集为平台格式，结束节点默认输出平台格式数据集。
  - 开始节点输入的数据集为非平台格式（Alpaca格式/ShareGPT格式）数据，结束节点默认输出对应同类型非平台格式数据。
  - 结束节点也可自行选择和开始节点不同格式数据集作为最终输出的数据集格式。
- 参数配置样例：

  无。
- 转换样例：

  输入节点处理前数据集格式：任意格式数据集。

  输出节点处理后输出数据集格式：任意格式数据集。

WORD 内容提取

- 适用的文件格式：“文档 > docx”。
- 各参数说明：

  待提取内容类型：从Word文档中提取文本，保留原文档的标题和正文等结构，不保留图片、公式、页眉、页脚，不支持嵌套表格提取。
- 参数配置样例：

  不需要配置参数，默认保留原文档的目录、标题和正文等结构，不保留图片、表格、公式、页眉、页脚。
- 提取样例：

  本地导入：{"fileName":"JAVA从入门到精通.docx","original_path": "Local Import","text":"JAVA是一种跨平台....."}

  OBS导入：{"fileName":"JAVA从入门到精通.docx","original_path": "nlp_data/word/JAVA从入门到精通.docx","text":"JAVA是一种跨平台....."}

AI Gallery: {"fileName":"JAVA从入门到精通.docx","original_path": "Gallery Subscription","text":"JAVA是一种跨平台....."}

CSV 内容提取

- 适用的数据集类型：“文本 > 单轮问答、单轮问答（人设）、问答排序”。
- 各参数说明：

  待提取内容类型：从CSV文件中读取所有文本内容，并按该文件内容类型模板KEY值生成匹配的JSON格式数据。
- 参数配置样例：

  不需要配置参数。
- 提取样例：

  如果提取CSV样式如："你好，请介绍自己，我是盘古大模型"，则提取内容输出为：｛"context":"你好，请介绍自己","target":"我是盘古大模型"｝

PDF 内容提取

- 适用的数据集类型：“文档 > pdf”。
- 各参数说明：

  待提取内容类型：默认保留文本、表格、公式和标题，支持选择需要保存的类型，未选择的类型将去除。

  精细化内容提取：是否支持版面分析完识别是图片的内容再次进行版面分析提取。

  表格提取可选格式：默认Latex，支持将表格转为Markdown格式。
- 参数配置样例：

- 提取样例：

  本地导入：{"fileName":"JAVA从入门到精通.pdf","original_path": "Local Import","text":"JAVA是一种跨平台....."}。

  OBS导入：{"fileName":"JAVA从入门到精通.pdf","original_path": "nlp_data/pdf/JAVA从入门到精通.pdf","text":"JAVA是一种跨平台....."}。

  AI Gallery：{"fileName":"JAVA从入门到精通.pdf","original_path": "Gallery Subscription","text":"JAVA是一种跨平台....."}。
- 算子限制：

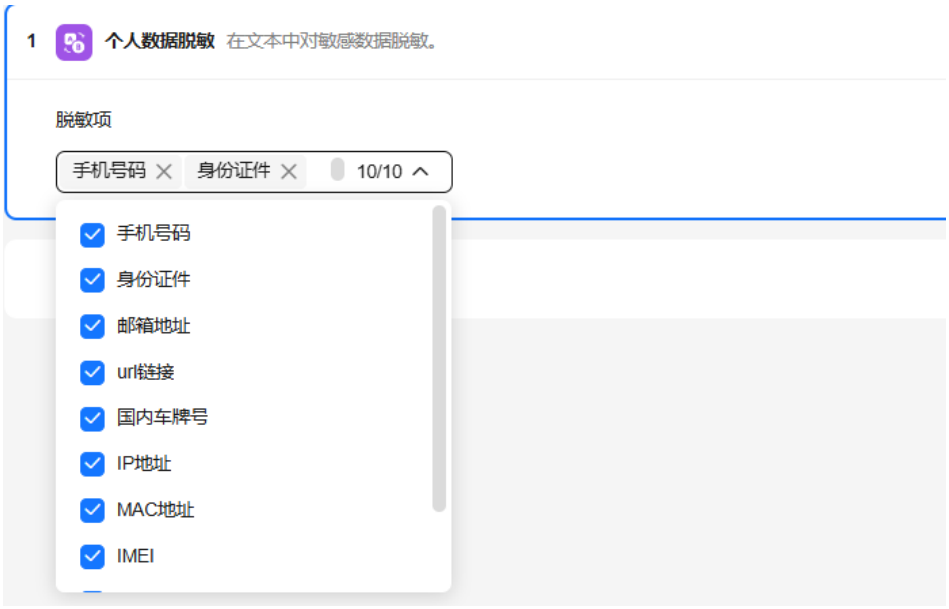
  pdf内容提取处理大规模数据时，运行时间超过24小时会中断，建议拆分后再执行。

个人数据脱敏

- 适用的数据集类型：“文本类”。
- 各参数说明：

待转换内容类型：对文本中的手机号码、身份证件、邮箱地址、URL链接、国内车牌号、IP地址、MAC地址、IMEI、护照、车架号等个人敏感信息进行数据脱敏，默认全部勾选，也可以选择部分。

- 参数配置样例：



- 转换样例：

  精炼前：“数据来自www.test.com”。

  精炼后：“数据来自*******”。

中文简繁互转

- 适用的数据集类型：“文本类”。
- 各参数说明：

  待转换内容类型：支持中文简体和中文繁体进行转换，过滤粒度为字符，默认转换方式为繁体转简体。
- 参数配置样例：

- 转换样例：

符号标准化

- 适用的数据集类型：“文本类”。
- 各参数说明：

待转换内容类型：支持对文本中携带的非标准化符号进行标准化、统一化转换，待标准化符号有空格、全角符号、标点符号、数字符号，默认全部勾选，过滤粒度为字符。

- 参数配置样例：

- 转换样例：根据映射表进行符号识别并映射。
  精炼前: {"fileName":"文本1.txt","text":"测试②①③非标准"}
  精炼后: {"fileName":"文本1.txt","text":"测试2.1.3.非标准"}

自定义正则替换

- 适用的数据集类型：“文本类”。
- 各参数说明：
  待转换内容类型：数据条目不变下，使用自定义正则表达式替换文本内容。
- 参数配置样例：

- 转换样例：
  精炼前: {"text":"这是aeiou正文内容aeiou测试aeiou。"}。
  精炼后: {"text":"这是11111正文内容11111测试11111。"}。

日期时间格式转换

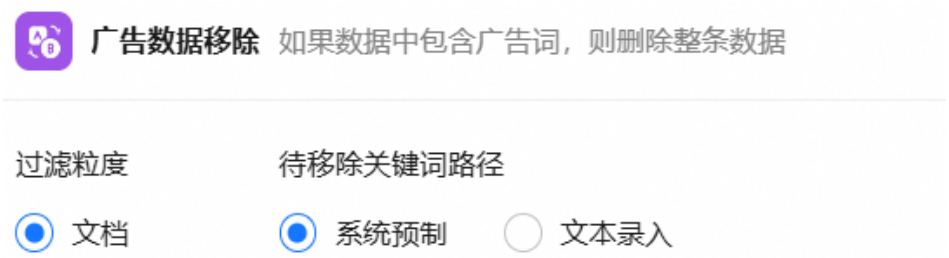
- 适用的数据集类型：“文本类”。
- 各参数说明：
  待转换内容类型：自动识别日期、时间、星期，同时根据选择的格式进行统一转换。转换类型包括日期格式、时间格式、星期格式，默认全都勾选，也支持选择部分进行转换。
- 参数配置样例：

- 转换样例：
  精炼前: {"text":"今天是2025年3月3号，周一，早上雨真大。"}。
  精炼后: {"text":"今天是2025-03-03 00:00:00，星期一，早上雨真大。"}。

广告数据移除

- 适用的数据集类型：“文本类”。
- 各参数说明：

  待过滤内容类型：按照句子的过滤粒度，删除文本中包含广告数据的句子。
- 参数配置样例：



- 过滤样例：

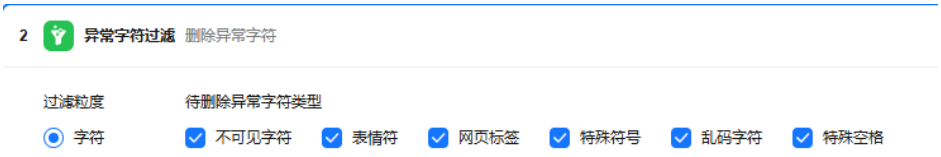
  精炼前: {"text": "※具体优惠信息！ ※购买我们的产品，享受高达50%的折扣！单击链接获取低价:https://example.com不要错过这个机会，赶快行动吧！ "}。

  精炼后: {"text":""}。

异常字符过滤

- 适用的数据集类型：“文本类”。
- 各参数说明：

  待过滤内容类型：查找数据集每一条数据中携带的异常字符，并将异常字符替换为空值，数据条目不变。异常字符过滤类型包括不可见字符、表情符、网页标签、特殊符号、乱码字符、特殊空格，默认全都勾选，也支持选择部分进行过滤。
- 参数配置样例：



- 过滤样例：

  精炼前: {"text":"测试异常●▌◆。<style></style>哈哈。限时特惠!☺"}。

  精炼后: {"text":"测试异常 。哈哈。 限时特惠!"}。

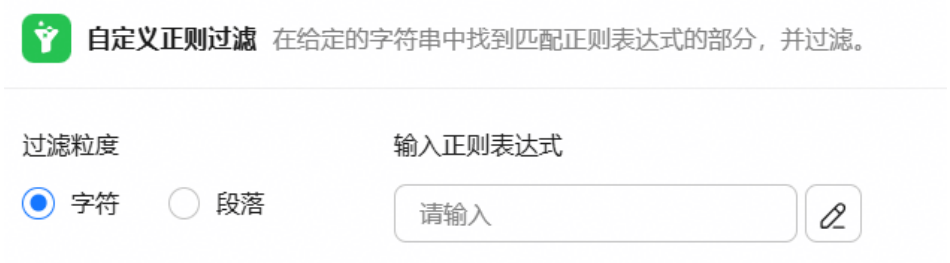
自定义正则过滤

- 适用的数据集类型：“文本类”。
- 各参数说明：

  待过滤内容类型：按自定义正则表达式进行匹配过滤，过滤粒度支持按字符、段落进行过滤，默认勾选字符。

  输入正则表达式：自定义正则过滤所需要的正则表达式。

  保留匹配样本：当待过滤内容类型为段落时展示，默认为否。
- 参数配置样例：



- 过滤样例：

  例如过滤掉参考文献之后的内容。

  精炼前: {"text":"这是正文内容。参考文献[1]作者1，文章1，期刊1，2021.[2] 作者2，文章2，期刊2，2022."}。

  精炼后: {"text":"这是正文内容。"}。

自定义关键词过滤

- 适用的数据集类型：“文本类”。
- 各参数说明：

  待过滤内容类型：过滤粒度支持按字符、段落、文档进行过滤，默认勾选字符。
  待删除的关键词路径支持从obs中导入关键词，以及文本录入。
- 参数配置样例：

- 过滤样例：

  例如按关键词测试进行过滤。

  精炼前: {"text":"关键词测试这是一条测试数据。"}。

  精炼后: {"text":"关键词这是一条数据。"}。

段落结尾不完整句子过滤

- 适用的数据集类型：“文本类”。
- 各参数说明：

  待过滤内容类型：按照句子的过滤粒度，自动识别段落结尾处的内容是否完整，如果不完整，则删除。
- 参数配置样例：

- 过滤样例：

  精炼前:"JAVA是一种面向对象的程序设计语言。使用JAVA语言。"

  精炼后:“JAVA是一种面向对象的程序设计语言。”

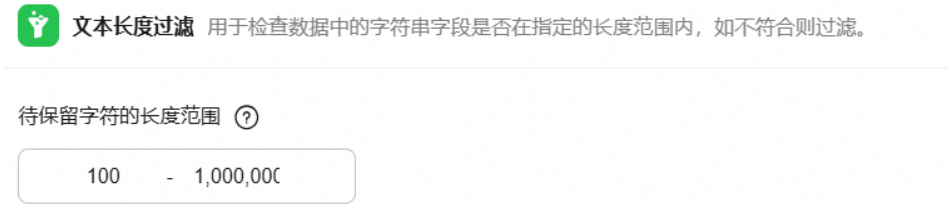
敏感词过滤

- 适用的数据集类型：“文本类”。
- 各参数说明：
  待过滤内容类型：对文本中涉及黄色、暴力、政治等敏感数据进行自动检测和过滤，需要预置敏感词。过滤粒度支持按字符、段落、文档进行过滤，默认勾选字符
- 参数配置样例：

- 过滤样例：
  精炼前: {"text":"嫖客啊fuck测试"}。
  精炼后: {"text":"啊测试"}。

文本长度过滤

- 适用的数据集类型：“文本类”。
- 各参数说明：
  待过滤内容类型：按照设置的文本长度，保留长度范围内的数据。默认待保留字符的长度范围为100-1000字符，支持修改，最小值为1。
- 参数配置样例：



- 过滤样例：
  精炼前：{"text": " 测试长度"}
  精炼后：{"text":""}

N-gram 特征过滤

- 适用的数据集类型：“文本类”。
- 各参数说明：
  组词数量N默认值为2，gram重复比率最大值默认值为0.15，计算阈值超过gram重复比率最大值的将会被过滤，支持修改。
- 参数配置样例：

- 过滤样例：

  精炼前: {"text":"醒醒，今天星期天，今天不上班，今天放假，明天星期一，明天上班。"}。

  精炼后: {"text":""}。

  注：该算子分词效果仅对中文数据有效，对英文数据无效果。

句子特征过滤

- 适用的数据集类型：“文本类”。
- 各参数说明：

  待过滤内容类型：按照文档过滤粒度，根据待保留的平均句长进行过滤，如果不满足则过滤。待保留的平均句长默认值为大于等于10个字符，支持修改，最小值为1。
- 参数配置样例：

- 过滤样例：

  精炼前: {"text":"在一个小村庄里，有一个传说。传说中，每当满月之夜，村庄的森林里会出现一只神秘的狐狸。"}。

  精炼后: {"text":""}。

违禁文本检测

- 适用的数据集类型：“问答排序、单轮问答、单轮问答（带人设）jsonl”。
- 各参数说明：选“是”则是过滤算子，选否则不过滤。
- 参数配置样例：

- 过滤样例：

  打标前：

  {"text": "QQ销售股民数据，你有吗"}

  打标后：

  {"text":"QQ销售股民数据，你有吗","text_ban_moderation":{"suggestion":"block","details":

{"confidence":1.0,"label":"violation_info","risk_level":2,"segments":[{"segment":"qq销售股民数据"},{"segment":"股民数据"},{"segment":"股民数据&销售"},{"segment":"销售股民数据"}],"suggestion":"block"}}

suggestion字段表示对文件检测是否通过的结果，pass代表审核通过，没有问题；review代表需要人工复核，您可以按照您的审核策略选择放通还是拦截；block代表待审文件存在问题。

个人隐私识别

- 适用的数据集类型：“问答排序、单轮问答、单轮问答（带人设）jsonl”。
- 各参数说明：选“是”则是过滤算子，选否则不过滤。
- 参数配置样例：

- 过滤样例：

  打标前：

  {"text": "你保存一下我的MAC地址：20-6E-D4-88-F3-98"}

  打标后：

{"text":"你保存一下我的MAC地址：20-6E-D4-88-F3-98","text_pii_moderation":{"suggestion":"block","details":[{"start":33,"end":50,"length":17,"data":"20-6E-D4-88-F3-98","category":"MAC_ADDRESS"}]}}

  suggestion字段表示对文件检测是否通过的结果，pass代表审核通过无相应的问题；review代表需要人工复核，您可以按照您的审核策略选择放通还是拦截；block代表待审文件存在问题。

垃圾内容文本检测

- 适用的数据集类型：“问答排序、单轮问答、单轮问答（带人设）jsonl”。
- 各参数说明：选“是”则是过滤算子，选否则不过滤。
- 参数配置样例：

- 过滤样例：

  打标前：

  {"text": "【开远假证848777596_qq合肥假证uhc0tm】什么意思_英语开远假证848777596_qq合肥假证uhc0tm的翻译_音标_读音_用法_例句_在线翻译_有道词典"}

打标后：

{"text":"【开远假证848777596_qq合肥假证uhc0tm】什么意思_英语开远假证848777596_qq合肥假证uhc0tm的翻译_音标_读音_用法_例句_在线翻译_有道词典","text_spam_moderation":{"details":[{"confidence":1.0,"label":"abuse","risk_level":2,"segments":[{"segment":"tm的"}],"suggestion":"block"}],"suggestion":"block"}}

suggestion字段表示对文件检测是否通过的结果，pass代表审核通过无相应的问题；review代表需要人工复核，您可以按照您的审核策略选择放通还是拦截；block代表待审文件存在问题。

广告文本检测

- 适用的数据集类型：“问答排序、单轮问答、单轮问答（带人设）jsonl”。
- 各参数说明：选“是”则是过滤算子，选否则不过滤。
- 参数配置样例：

- 过滤样例：

  打标前

  {"context": "清仓大甩卖，全场只要2元", "target": "价格好便宜"}

  打标后

  {"context":"清仓大甩卖，全场只要2元","target":"价格好便宜","text_ad_moderation":{"details":[],"suggestion":"pass"}}

  suggestion字段表示对文件检测是否通过的结果，pass代表审核通过无相应的问题；review代表需要人工复核，您可以按照您的审核策略选择放通还是拦截；block代表待审文件存在问题。

色情文本检测

- 适用的数据集类型：“问答排序、单轮问答、单轮问答（带人设）jsonl”。
- 各参数说明：选“是”则是过滤算子，选否则不过滤。
- 参数配置样例：

- 过滤样例：

  打标前：

{"text": "狼友黄站导航，现在就来快乐爆操，让的生活充满色情和刺激，还等"}

打标后：

{"text":"狼友黄站导航，现在就来快乐爆操，让的生活充满色情和刺激，还等","text_porn_moderation":{"suggestion":"block","details":"[{'confidence': 1.0, 'label': 'porn_violence', 'risk_level': 2, 'segments': [{'segment': '爆操'}, {'segment': '狼友黄站导航'}], 'suggestion': 'block'}]"}}

suggestion字段表示对文件检测是否通过的结果，pass代表审核通过无相应的问题；review代表需要人工复核，您可以按照您的审核策略选择放通还是拦截；block代表待审文件存在问题。

辱骂文本检测

- 适用的数据集类型：“问答排序、单轮问答、单轮问答（带人设）jsonl”。
- 各参数说明：选“是”则是过滤算子，选否则不过滤。
- 参数配置样例：

- 过滤样例：

  打标前：

  {"text": "谁要和你一起死要死你自己死"}

  打标后:

  {"text":"谁要和你一起死要死你自己死","text_abuse_moderation":{"details":[{"confidence":0.9998,"label":"abuse","risk_level":2,"segments":[],"suggestion":"block"}],"suggestion":"block"}}

  suggestion字段表示对文件检测是否通过的结果，pass代表审核通过无相应的问题；review代表需要人工复核，您可以按照您的审核策略选择放通还是拦截；block代表待审文件存在问题。

涉政文本检测

- 适用的数据集类型：“问答排序、单轮问答、单轮问答（带人设）jsonl”。
- 各参数说明：选“是”则是过滤算子，选否则不过滤。
- 参数配置样例：

- 过滤样例：

  打标前:

  {"text": "但中共当局对这些网络质疑声音从来不屑于解释，而是直接封杀"}

  打标后:

  {"text":"但中共当局对这些网络质疑声音从来不屑于解释，而是直接封杀","text_pollnfo_moderation":{"suggestion":"block","details":"[{'confidence': 1.0, 'label': 'politics', 'risk_level': 3, 'segments': [{'segment': '中共当局'}], 'suggestion': 'block'}]"}}

  suggestion字段表示对文件检测是否通过的结果，pass代表审核通过无相应的问题；review代表需要人工复核，您可以按照您的审核策略选择放通还是拦截；block代表待审文件存在问题。

预训练文本分类

- 适用的数据集类型：“文档、预训练文本”。
- 各参数说明：

  待打标内容类型：针对预训练文本进行内容分类，例如新闻、教育、健康等类别，支持分析语种包括：中文、英文，默认中文。
- 参数配置样例：

- 打标样例：

  {"fileName":"新闻打标测试.docx","text":" 本报北京3月3日电(记者徐佩玉)中国人民银行发布的今年1月份金融市场运行情况显示，1月份，我国债券市场共发行各类债券51027.5亿元。其中，国债发行10185.0亿元,地方政府债券发行5575.7亿元，金融债券发行7042.1亿元，公司信用类债券发行12791.7亿元，信贷资产支持证券发行27.3亿元，同业存单发行15147.8亿元。\n截至1月末,我国债券市场托管余额178.2万亿元。其中，银行间市场托管余额156.9万亿元，交易所市场托管余额21.3万亿元。\n在债券市场对外开放方面,截至1月末，境外机构在中国债券市场的托管余额4.2万亿元，占中国债券市场托管余额的比重为2.3%。其中，境外机构在银行间债券市场的债券托管余额4.1万亿元：分券种看，境外机构持有国债2.0万亿元、占比48.8%，同业存单1.1万亿元、占比25.8%，政策性银行债券0.9万亿元、占20.8%。\n","pre_classification":"经济"}

数据生成

- 适用的数据集类型：“单轮问答、单轮问答（带人设）”。
- 各参数说明：

  生成场景：对于单轮问答、单轮问答人设的输入数据，可以对数据进行一系列合成操作，如问题生成回答、问答对改写等，选择对应使用场景即可一键生成。

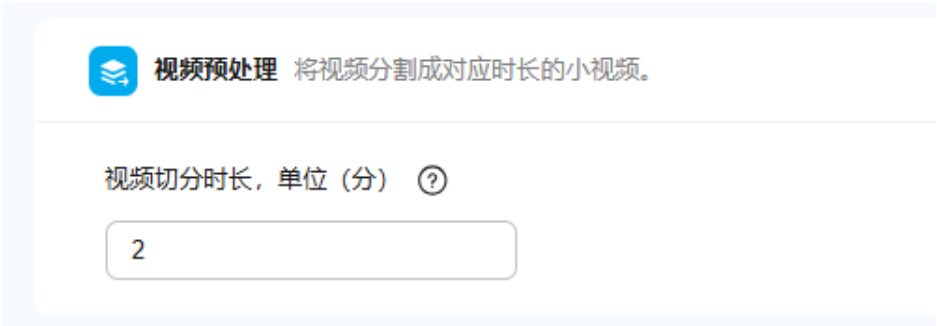
  模型：选择需要用于数据生成的模型
- 算子功能描述：支持从单一样本生成相似问答、为问答注入特定人设角色，并可一键调整问答难度，实现数据的规模化定制合成。

- 参数配置样例：

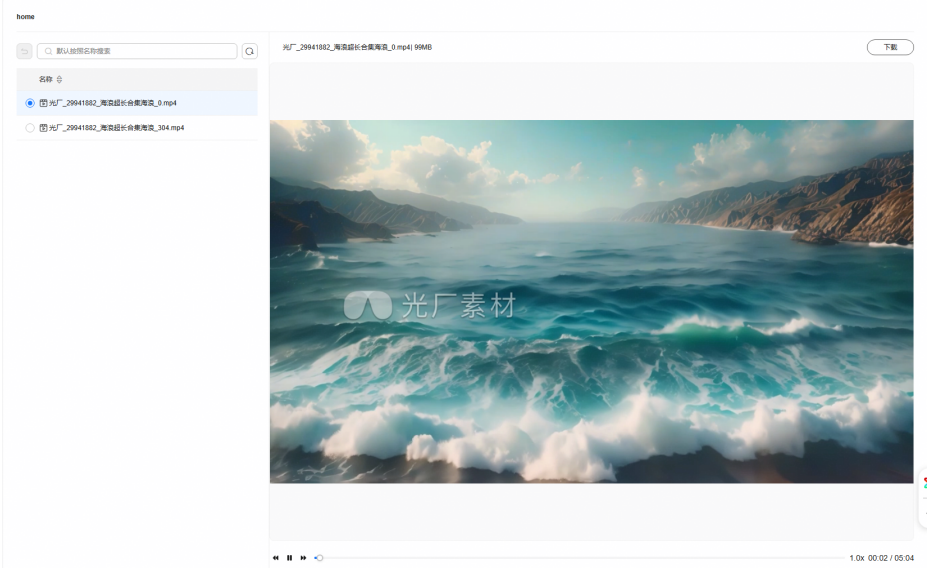
视频时长切分

- 适用的文件格式：“视频>mp4 / avi”。
- 各参数说明：

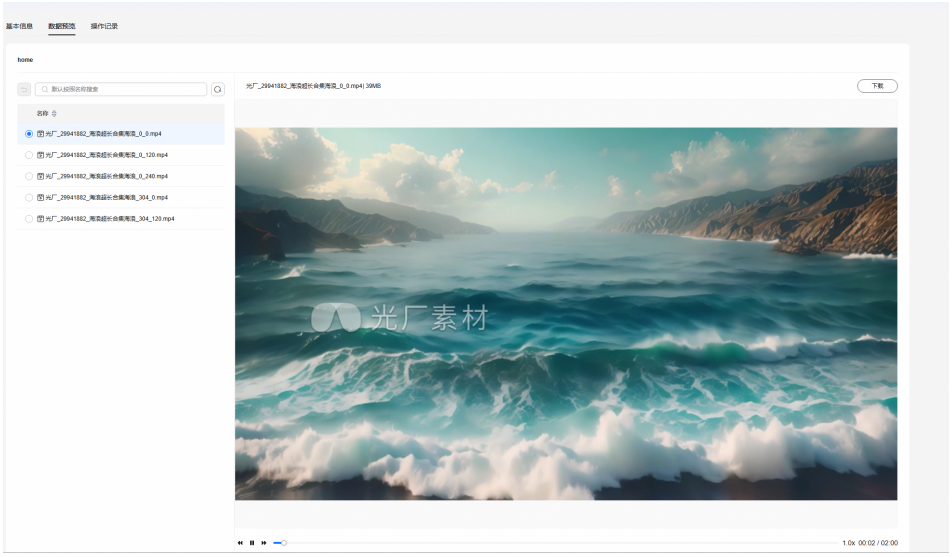
  视频切分时长：配置该参数可以确定切分后的视频时长，范围是1-5分钟。如果源视频时长不满足需要切分的条件，则保留源视频。
- 算子功能描述：将源视频切分成固定时长的小视频，固定时长可配置，范围为1-5分钟。先进行视频切分将视频长度减小再使用镜头切分会提高算子效率。
- 使用场景：
  - 可处理情况
    - 视频时长大于1min。
  - 暂无法解决情况
    - 视频时长小于1min。
- 参数配置样例：



- 视频切分后时长对比
  - 视频切分前：



- 视频切分后：



镜头拆分

- 适用的文件格式：“视频>mp4 / avi”。
- 各参数说明：

  需要拆分的视频：筛选出分辨率、时长、帧率同时满足筛选标准的视频进行镜头拆分。

  视频拆分后规格：单视频切片最大时长支持自定义；如果首轮拆分切片时长超过设定值，则会进一步做拆分，最终拆分结果均小于等于设定阈值 。
- 使用场景：
  - 可处理情况
    - 有显著场景变换，包含直接切换或者淡入淡出。
  - 暂无法解决情况

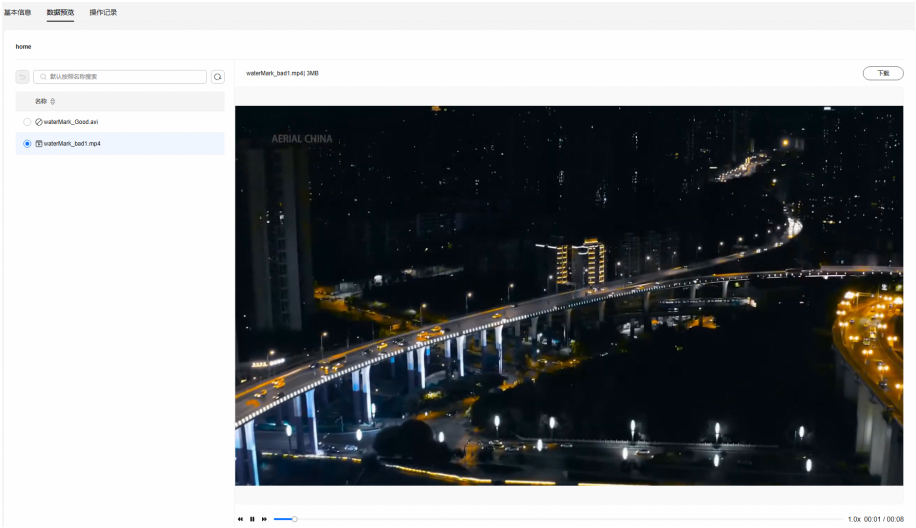
- 同一场景拍摄内容跳变但内容相似度高。

- 参数配置样例：

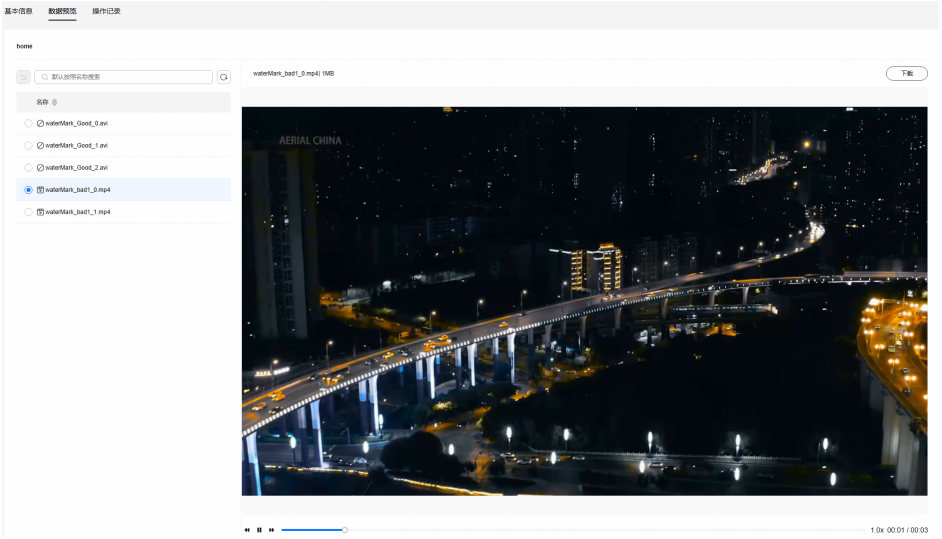
拆分样例：设置单视频切片最大时长3秒：

镜头拆分前后对比：

- 拆分前：



- 拆分后：



视频裁剪

- 适用的文件格式：“视频>mp4 / avi”。
- 各参数说明：

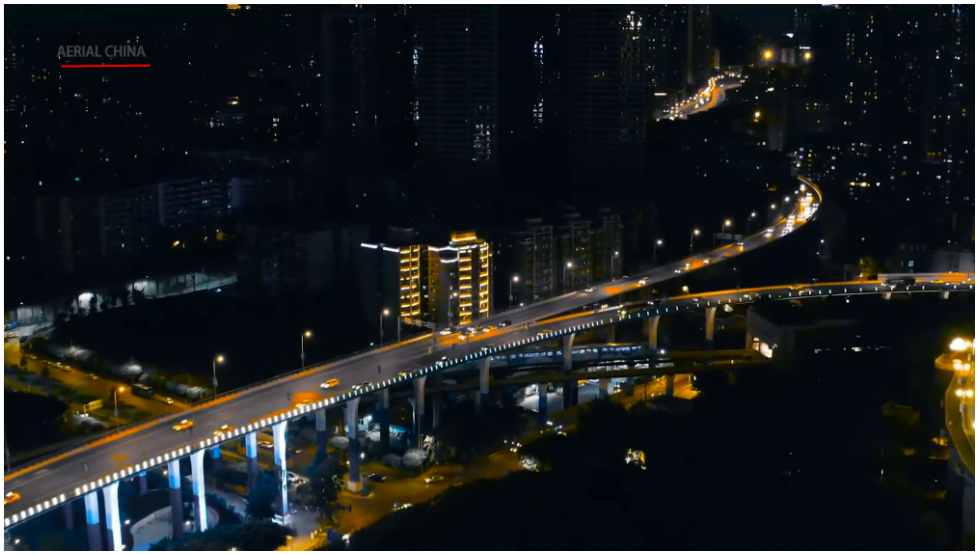
  裁剪项：自定义选择裁剪项，裁剪掉视频中字幕/Logo/水印/边框/密集文字等无用信息。

  最大裁剪比例：裁剪视频面积/原始视频面积的值即裁剪面积占比，设置默认的裁剪比例，默认值为0.3。

  过裁剪保留：裁剪占比大于最大裁剪比例时，是否保留原视频。是则保留，否则过滤。
- 使用场景：
  - 可处理情况
    - 需要先执行字幕、logo、水印、边框、密集文字识别算子。
  - 暂无法解决情况
    - 未先执行字幕、logo、水印、边框、密集文字识别算子。
    - 裁剪后无法保留留存过小或者比例失衡的视频。
- 参数配置样例：

- 裁剪样例：

  裁剪前：带水印视频。



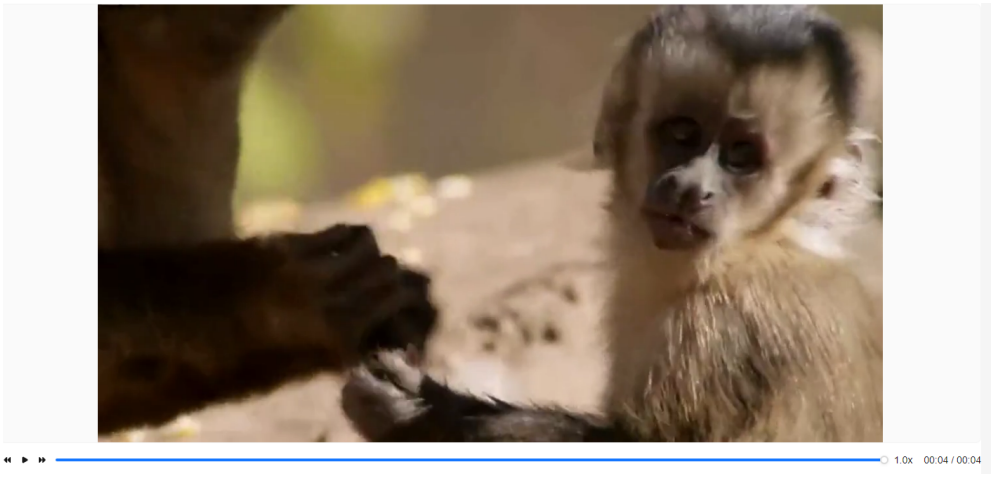
  裁剪后：上部带水印部分被裁剪，视频高度变低。

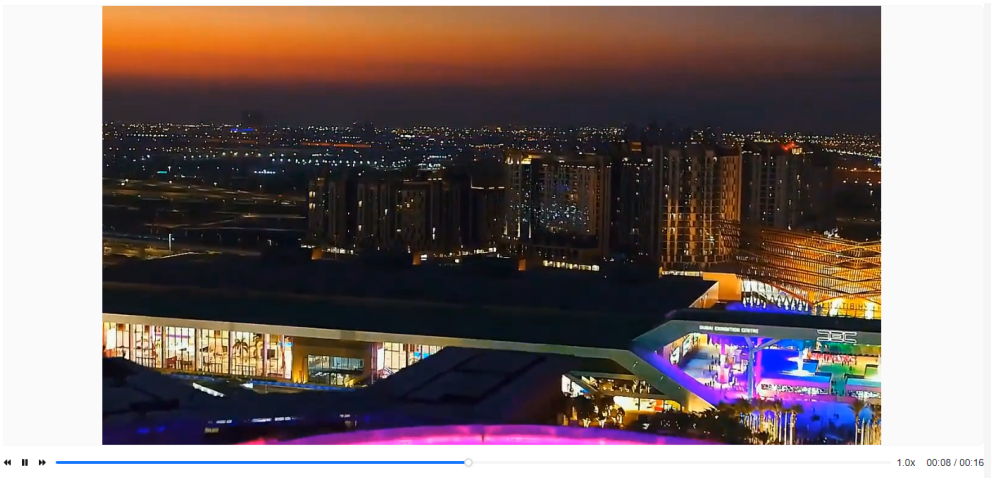
视频元数据过滤

- 适用的文件格式：“视频>mp4 / avi”。
- 各参数说明：
  待保留分辨率：自定义选择保留分辨率。不满足所选分辨率的视频将被过滤掉。
  待保留时长：默认值为3，小于“待保留时长”的视频将被过滤掉。
  待保留帧率：电影标准帧率为24或30FPS，小于“待保留帧率”的视频将被过滤掉。
- 参数配置样例：

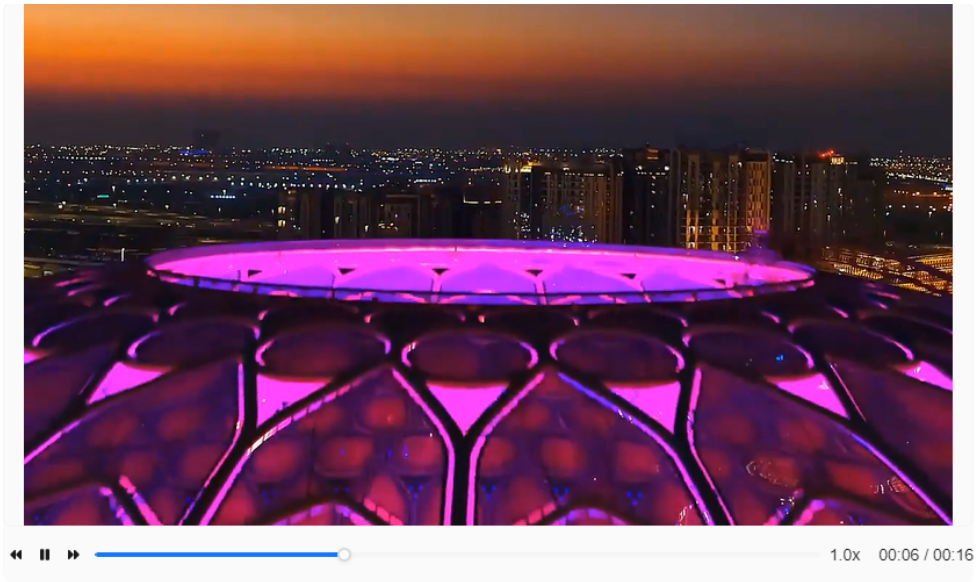
- 过滤样例：设置待保留时长大于等于10S：

过滤前：两个视频，一个时长是4S，一个时长是16S。





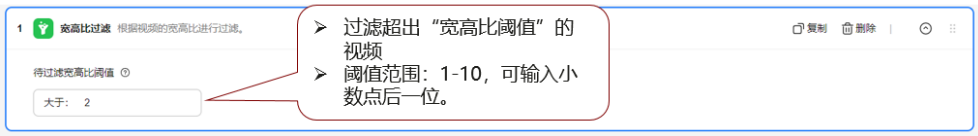
过滤后：只保留时长为16S的视频：



视频宽高比过滤

- 适用的文件格式：“视频>mp4 / avi”。
- 各参数说明：

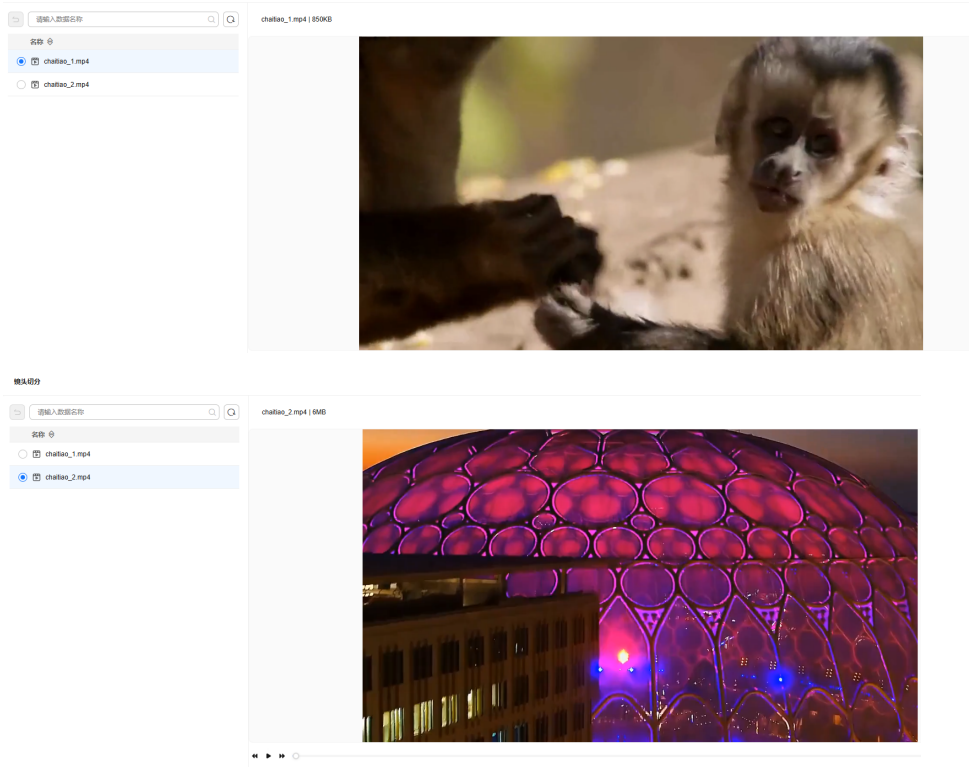
  待过滤宽高比阈值：超出“宽高比阈值”的视频将被过滤掉。阈值范围为(1, 10)，可输入小数点后一位。
- 参数配置样例：



- 过滤样例：

  原视频数据集：

  共有两个视频，第一个宽高比为1.77，第二个宽高比为1.79



设置宽高比阈值为1.78，经算子处理过后，仅保留宽高比为1.79的视频。

色情视频检测

- 适用的文件格式：“视频>mp4 / avi”。
- 算子说明：给色情视频内容打标签。
- 参数配置样例：
  不需要配置参数。

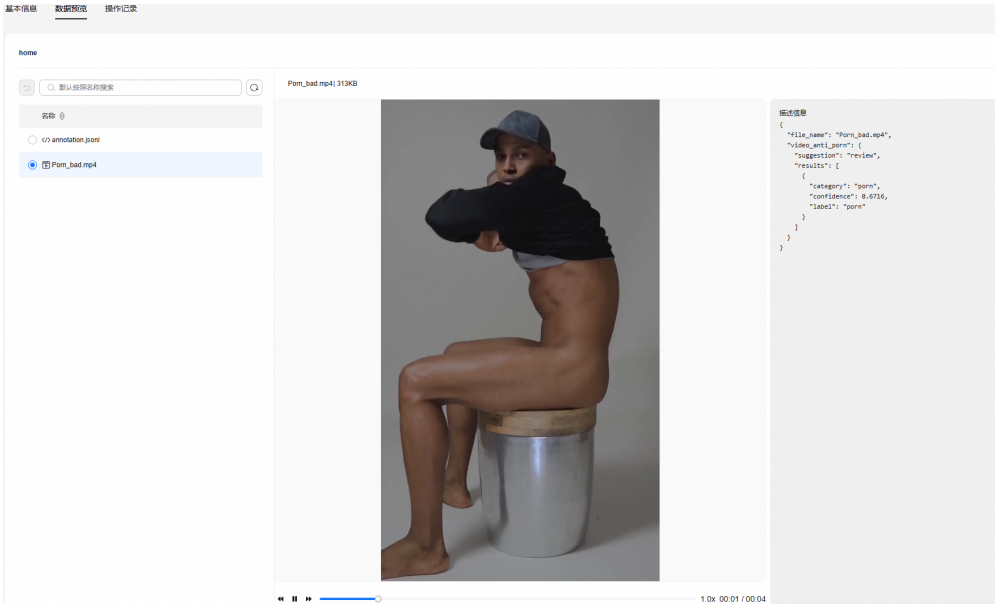
- 检测样例：

  检测结果以video_anti_porn对象存储在标注文件中。

  suggestion：对文件检测是否通过的结果，pass代表审核通过无相应的问题；review代表需要人工复核，您可以按照您的审核策略选择放通还是拦截；block代表待审文件存在问题。

  confidence：模型结果检测的置信度（注意这里的置信度代表模型给出建议的置信度）。如果suggestion为pass，则为零；如果suggestion为review/block，则为0-1。

  label：模型检测出的具体色情标签，如果未检测出则为空。



暴恐视频检测

- 适用的文件格式：“视频>mp4 / avi”。
- 算子说明：给暴恐视频内容打标签。
- 参数配置样例：

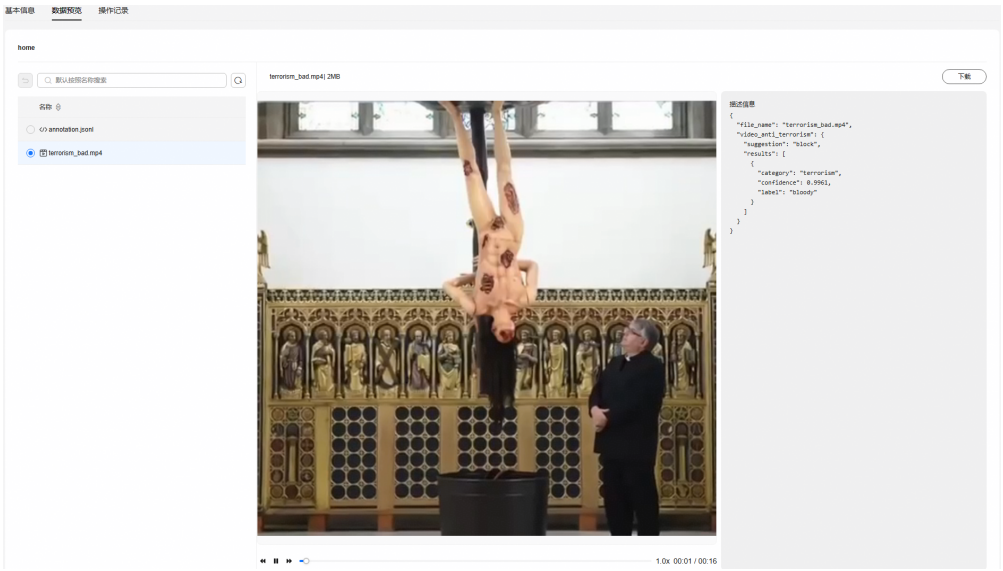
  不需要配置参数。

- 检测样例：检测结果以video_anti_terrorism对象存储在标注文件中。

  suggestion：对文件检测是否通过的结果，pass代表审核通过无相应的问题；review代表需要人工复核，您可以按照您的审核策略选择放通还是拦截；block代表待审文件存在问题。

  confidence：模型结果检测的置信度（注意这里的置信度代表模型给出建议的置信度）。如果suggestion为pass，则为零；如果suggestion为review/block，则为0-1。

  label：模型检测出的具体暴恐标签，如果未检测出则为空。



视频涉政检测

- 适用的文件格式：“视频>mp4 / avi”。
- 算子说明：

  给涉政视频内容打标签。
- 参数配置样例：

  不需要配置参数。
- 使用场景：

  主要检测国内政治人物、国外政治人物、国内负面政治领导人物、国外恐怖分子、国外的异端头目等，暂无法保证完全识别准确。
- 检测样例：

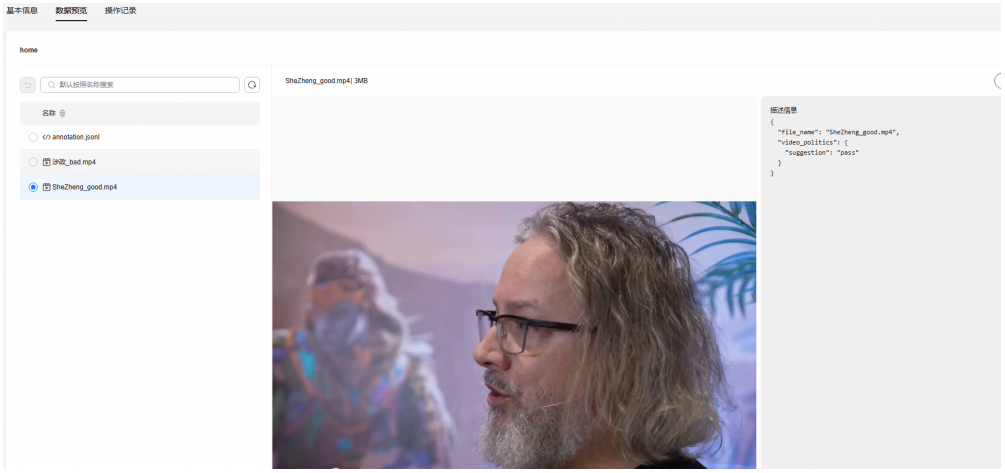
  检测结果以video_anti_politics对象存储在标注文件中。

  suggestion：对文件检测是否通过的结果，pass代表审核通过无相应的问题；review代表需要人工复核，您可以按照您的审核策略选择放通还是拦截；block代表待审文件存在问题。

  result：模型对文件检测的具体返回内容，包含suggestion、confidence、label三个子标签；可以一条或多条。

  confidence：模型结果检测的置信度（注意这里的置信度代表模型给出建议的置信度）。如果suggestion为pass，则为零；如果suggestion为review/block，则为0-1。

  label：模型检测出的具体涉政标签，如果未检测出则为空。



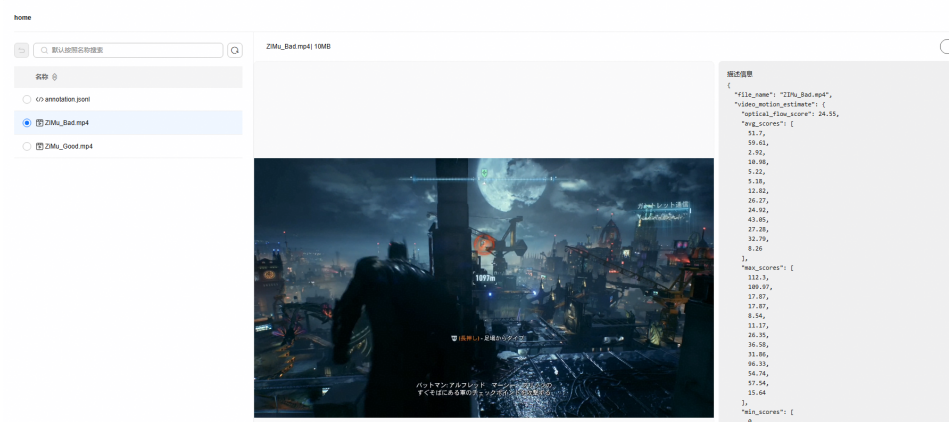
运动幅度评分

- 适用的文件格式：“视频>mp4 / avi”。
- 评分说明：

  识别运动幅度过快或过慢的视频，数值越大表示运动越快。运动幅度＞100光流可视为运动过快，运动幅度≤2光流可视为运动过慢。
- 使用场景：
  - 可处理情况
    - 画面运动幅度过大或过小，以及静止的画面可以识别。
  - 暂无法解决情况
    - 无法对快速/慢速占比小的部分进行识别。
- 参数配置样例：

- 评分样例：jsonl文件中显示运动幅度评分：

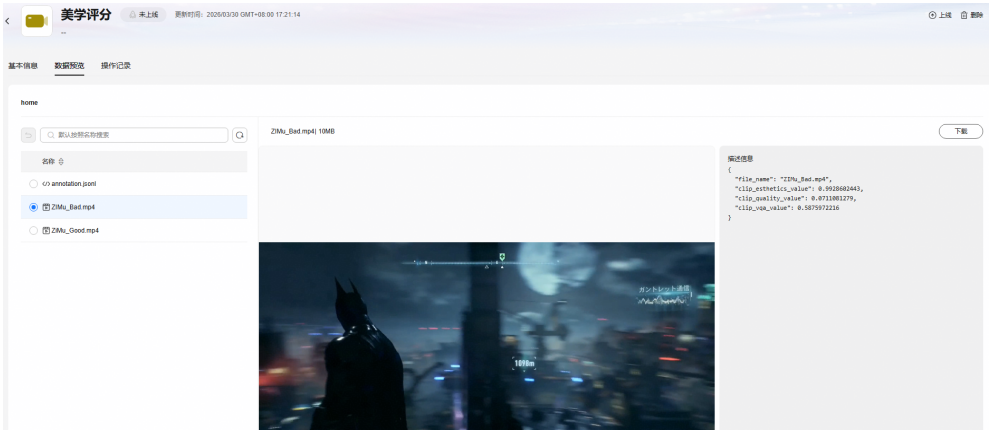
美学评分

- 适用的文件格式：“视频>mp4 / avi”。
- 评分说明：

  从内容（吸引人，清晰度）、构图（目标物位置良好）、颜色（有活力，令人愉悦）、光线（光线明显有对比度）、轨迹（连续、稳定）等维度评价视频美感得分。分值范围(0, 1)，数值越高美感越好，评分>0.95可视为视频美感较高的视频。
- 使用场景：
  - 可处理情况
    - 美学问题或质量比较明显的视频识别效果较好。
  - 暂无法解决情况
    - 无法处理像素游戏这种类型的视频。
    - 对水印不敏感。
- 参数配置样例：

- 评分样例：jsonl文件中显示美学评分：clip_esthetics_value：美学分

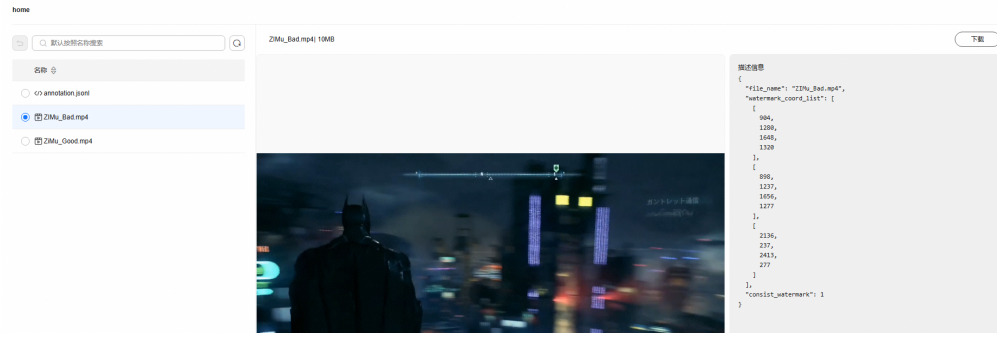
水印识别

- 适用的文件格式：“视频>mp4 / avi”。
- 算子说明：

  识别视频中是否包含水印。
- 参数配置样例：

  水印识别阈值：当水印识别可信度高于水印识别阈值时即判断存在水印，默认水印识别阈值为0.5。
- 参数配置样例：

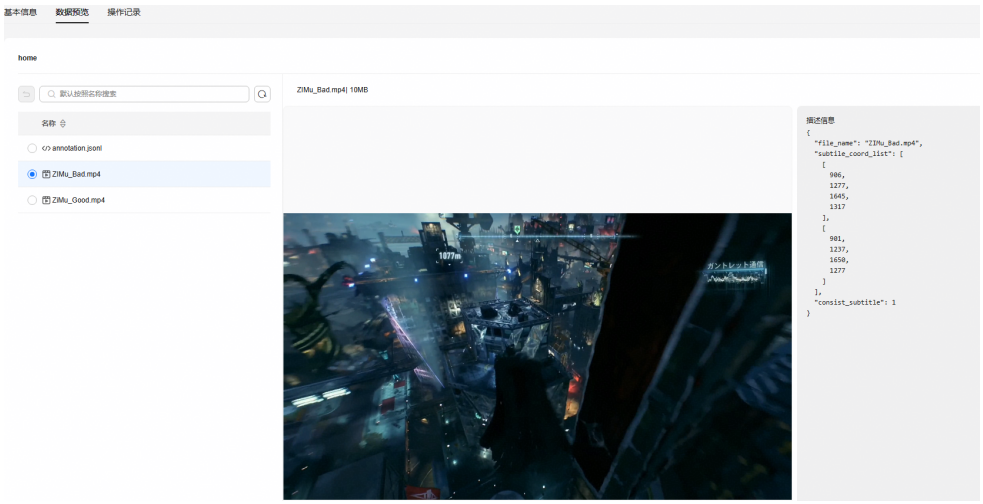
- 识别样例：jsonl文件中显示是否识别水印：consist_watermark值为1表示识别到水印，值为0表示未识别到水印。



字幕识别

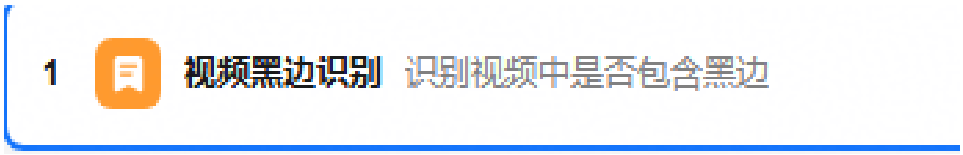
- 适用的文件格式：“视频>mp4 / avi”。
- 算子说明：
  识别视频中是否包含字幕。
- 参数配置样例：

- 识别样例：jsonl文件中显示是否识别字幕：consist_subtitle值为1表示识别到字幕，值为0表示未识别到字幕。

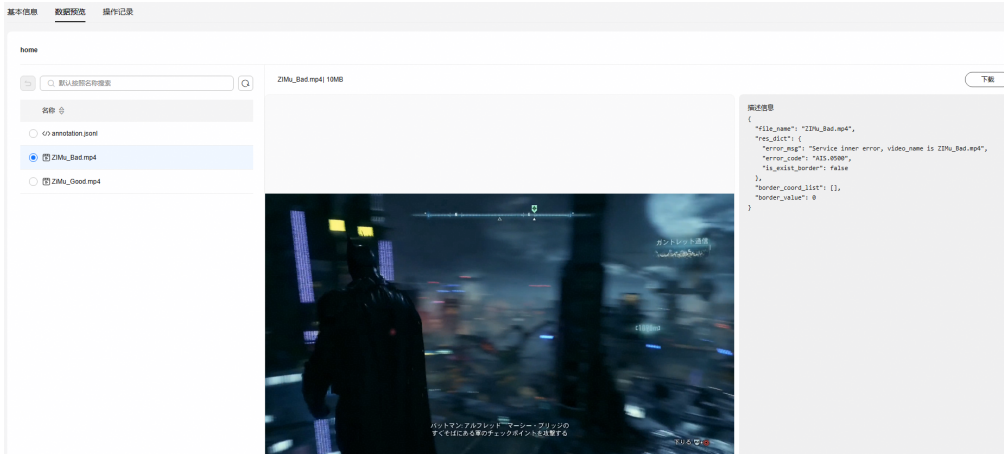


视频黑边识别

- 适用的文件格式：“视频>mp4 / avi”。
- 算子说明：
  识别视频中是否包含黑边。
- 使用场景：
  - 可处理情况
    - 只能处理视频的四个边，并且黑边的色差波动不大。
  - 暂无法解决情况
    - 无法处理不在四边，并且黑边内有其他字幕等色差变化的视频。
- 参数配置样例：



- 识别样例：border_value为1表示识别出黑边，值为0表示未识别出黑边

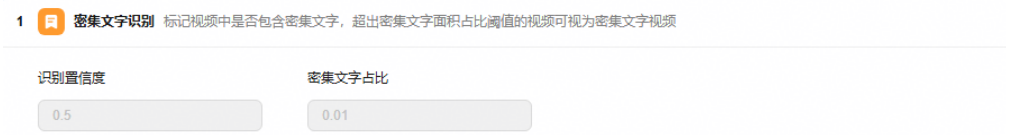


密集文字识别

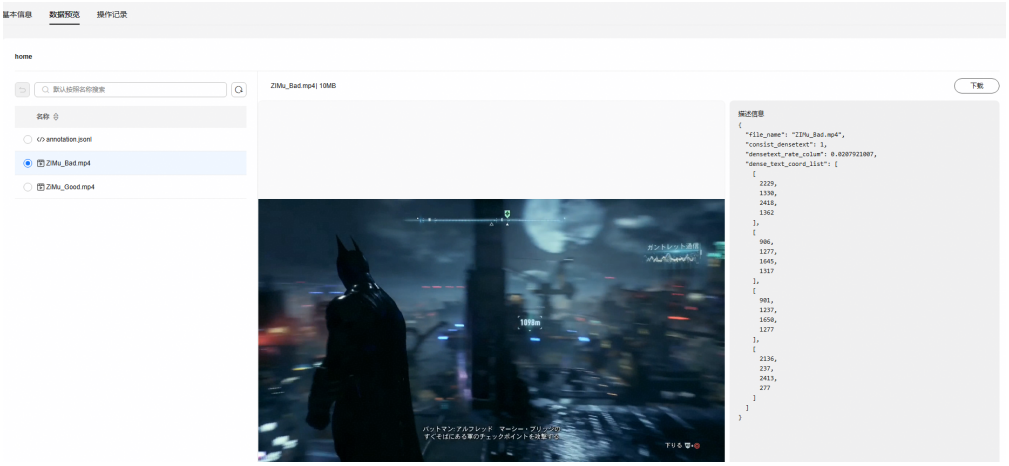
- 适用的文件格式：“视频>mp4 / avi”。
- 参数说明：

  密集文字面积占比：超出密集文字面积占比阈值的视频可视为密集文字视频，一般密集文字面积占比阈值为1%。

  置信度：当识别置信度超过设定阈值时，即可认定为包含密集文字的视频内容。默认情况下，识别置信度阈值设为 0.5。
- 参数配置样例：



- 识别样例：jsonl文件中显示是否识别密集文字：consist_densetext值为1表示识别到密集文字，值为0表示未识别到密集文字。



视频分类

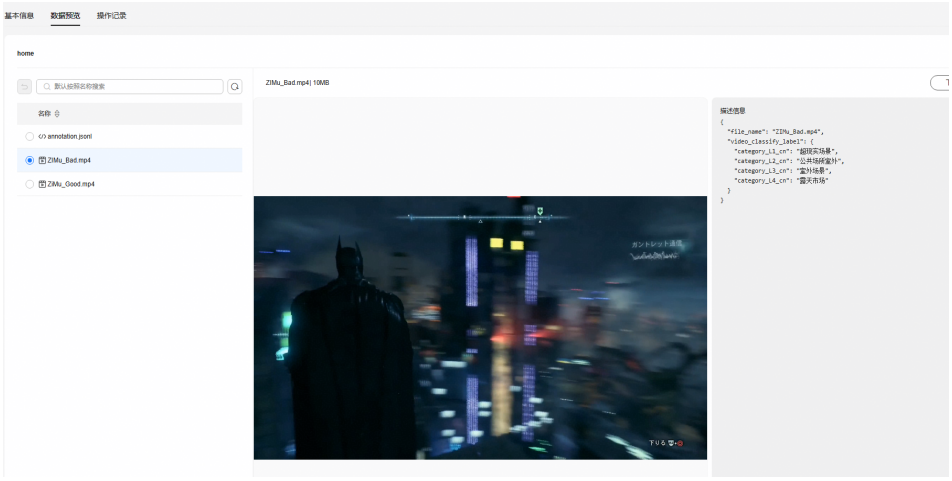
- 适用的文件格式：“视频>mp4 / avi”。
- 算子说明：

  自动对短视频内容进行分类，并生成相应的标签。
- 使用场景：
  - 可处理情况
    - 预设的类别可以进行分类。
  - 暂无法解决情况
    - 分类精度未作验证，只用来均匀采样。
    - 不支持非预设类别分类
- 参数配置样例：

  无需配置参数。
- 分类标注样例：

  描述信息中显示视频的各级分类：

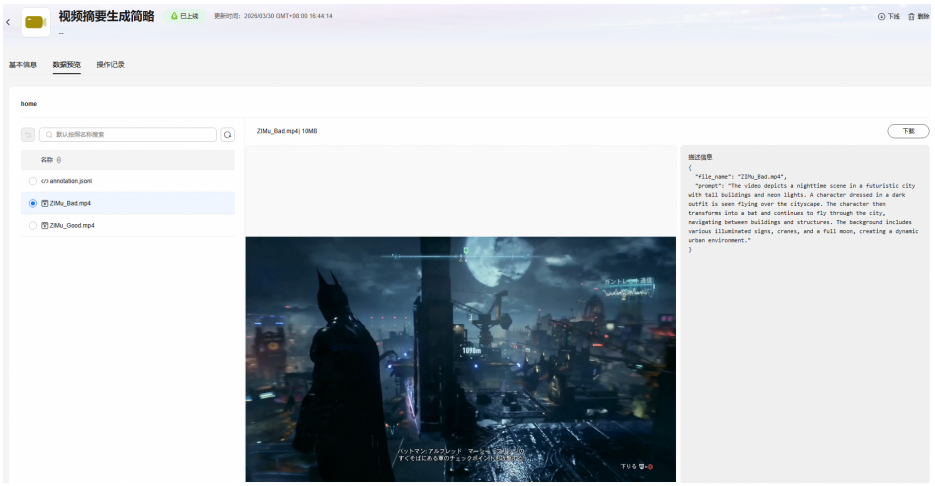
category_L1_cn：一级分类。
category_L2_cn：二级分类。
category_L3_cn：三级分类。
category_L4_cn：四级分类。



视频摘要生成（简略）

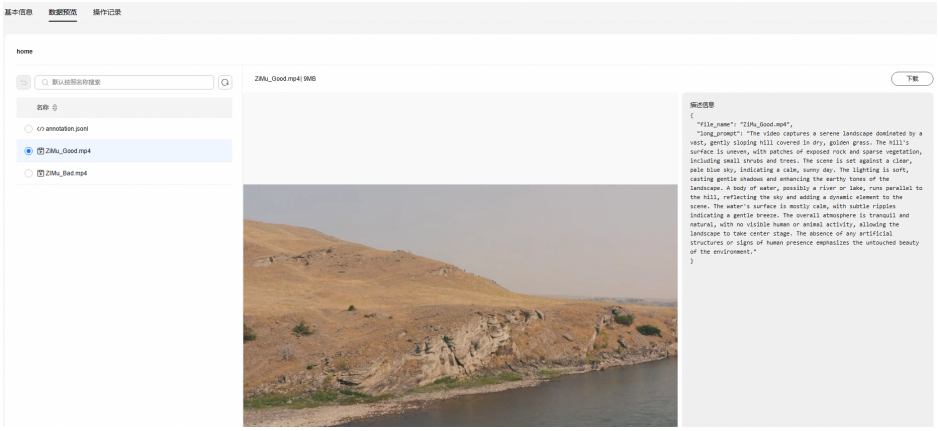
- 适用的文件格式：“视频>mp4 / avi”。
- 算子说明：
  通过对视频进行抽帧，通过模型推理生成简略的视频摘要描述。
- 使用场景：
  - 可处理情况
    - 所有视频都可以进行简短描述。
  - 暂无法解决情况
    - 无法指定描述方式。
    - 只能对视频的观感信息（场景、外观、行为）进行描述，无法理解视频深度内容（如新闻理解、内容解读、知名人物识别等），无法处理音频。
- 参数配置样例：
  无需参数配置。
- 打标样例：描述信息中prompt字段代表简略的视频摘要。

图 4-39 打标样例



视频摘要生成（详细）

- 适用的文件格式：“视频>mp4 / avi”。
- 算子说明：
  通过对视频进行抽帧，通过模型推理生成详细的视频摘要描述。
- 使用场景：
  - 可处理情况
    - 所有视频都可以进行描述。
  - 暂无法解决情况
    - 无法指定描述方式。
    - 非常详细的内容，如数量、动作细节等无法精确描述。
    - 只能对视频的观感信息（场景、外观、行为）进行描述，无法理解视频深度内容（如新闻理解、内容解读、知名人物识别等），无法处理音频。
- 参数配置样例：
  无需参数配置。
- 打标样例：描述信息中long_prompt字段代表详细的视频摘要

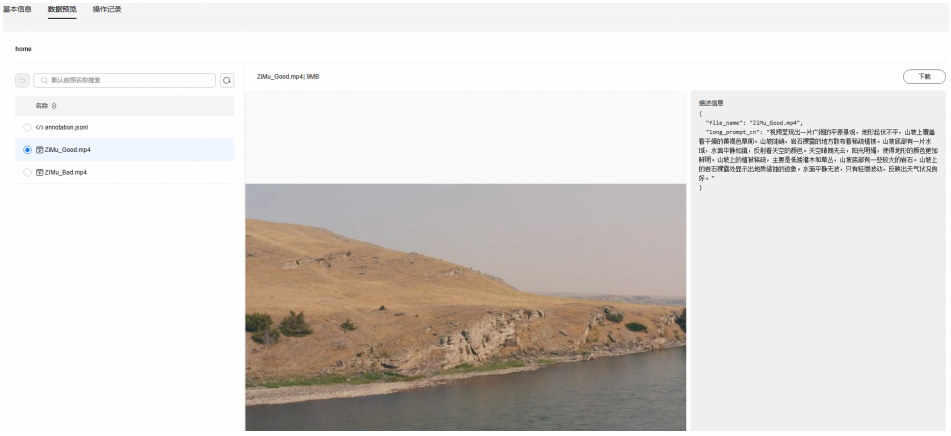


视频中文摘要生成（详细）

- 适用的文件格式：“视频>mp4 / avi”。
- 算子说明：

  通过对视频进行抽帧，通过模型推理生成详细的视频中文摘要描述。
- 使用场景：
  - 可处理情况
    - 所有视频都可以进行描述。
  - 暂无法解决情况
    - 无法指定描述方式。
    - 非常详细的内容，如数量、动作细节等无法精确描述。
    - 只能对视频的观感信息（场景、外观、行为）进行描述，无法理解视频深度内容（如新闻理解、内容解读、知名人物识别等），无法处理音频。
- 参数配置样例：

  无需参数配置。
- 打标样例：描述信息中long_prompt_cn字段代表详细的视频摘要。



姿态检测

- 适用的文件格式：“视频>mp4 / avi”。
- 算子说明：

  姿态检测算子从视频中抽取8帧图片，对每帧图片进行关键点打标与置信度计算，并计算符合过滤条件的图片数量，达到一定数量则表示视频中存在对应的人物数量。
- 使用场景：
  - 可处理情况
    - 可处理有人物脸部露出的视频。
  - 暂无法解决情况

    - 人物被部分遮挡会导致识别失败。
- 参数配置样例：

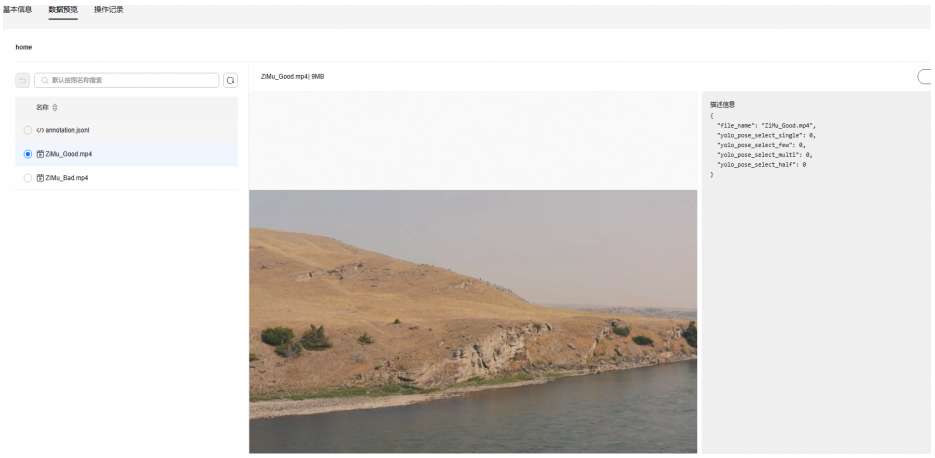
  无需参数配置。
- 打标样例：

  yolo_pose_select_single: 是否检测到了单个人的姿势，存在为1，否则为0。

  yolo_pose_select_few: 是否检测到了少量人（通常为2-4）的姿势，存在为1，否则为0。

  yolo_pose_select_multi: 是否检测到了多人（通常是4人或更多）的姿势，存在为1，否则为0。

  yolo_pose_select_half : 是否检测到了半个人的姿势，存在为1，否则为0。



镜头运动描述

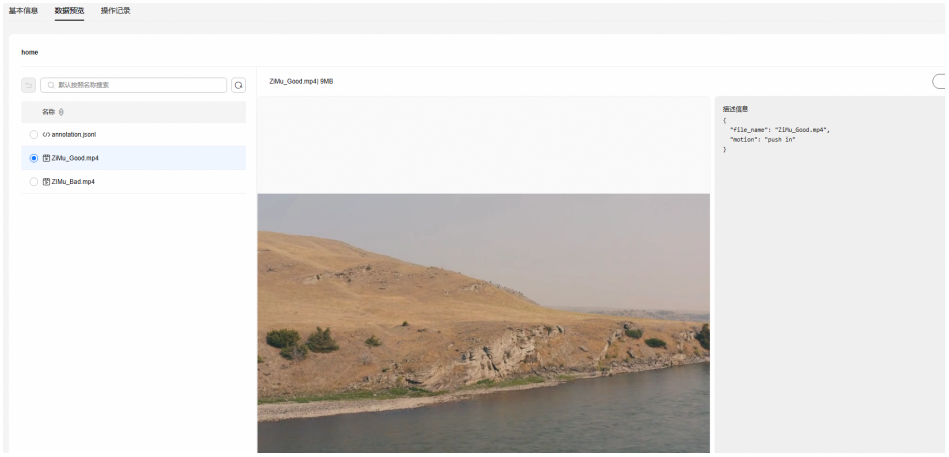
- 适用的文件格式：“视频>mp4 / avi”。
- 算子说明：

  模型通过对视频进行抽帧进行光流计算与推理，输出视频的镜头类型。
- 使用场景：
  - 可处理情况
    - 视频中运镜明确且不混乱。
  - 暂无法解决情况
    - 多种运镜组合或不明显会导致无法准确识别，只能识别预设的类别。
- 参数配置样例：

  无需参数配置。
- 打标样例：

  motion: 运镜的类型。

  标签范围为： { 0: 'static', 1: 'others', 2: 'pull out', 3: 'push in', 4: 'static' , 5: 'tracking', 6: 'orbit', 7: 'spin', 8: 'tilt up', 9: 'tilt down', 10: 'pan right', 11: 'pan left' ,12: 'tracking' }。



图文提取

- 适用的文件格式：

  tar+jsonl；所有图片保存为tar包。图片格式支持：jpg、jpeg、png、bmp。图片文本保存为一份jsonl文件，jsonl文件中图片名称必须要与tar包中的图片名称一致。

- 各参数说明：

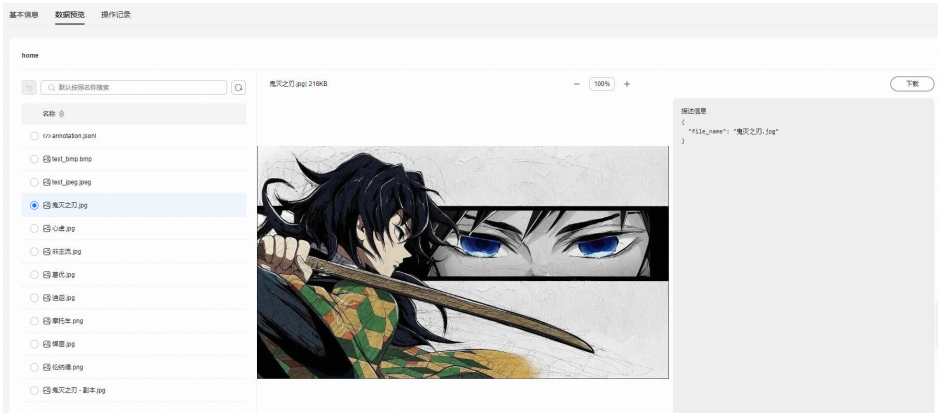
  待提取内容类型：提取图文压缩包中的JSON文本和图片；并对图片进行结构化解析。

- 参数配置样例：

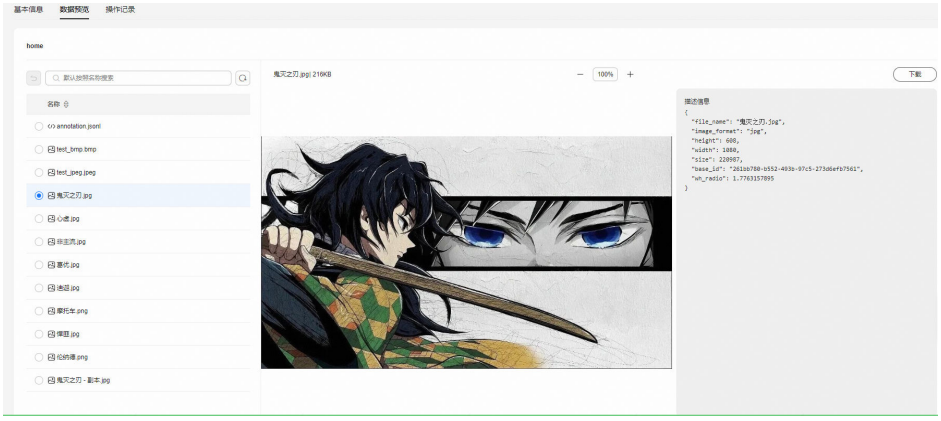
  不需要配置参数。

- 提取样例：

  精炼前：



  精炼后：



图片元数据过滤

- 适用的文件格式：

  jpg、jpeg、png、bmp。

  tar：所有图片保存为tar包。tar包含图片支持：jpg、jpeg、png、bmp图片类型。

- 各参数说明：

  待过滤内容类型：

  最小宽：宽低于此设置值，图片会被过滤。

  最小高：高低于此设置值，图片会被过滤。

  最小宽高比：图片宽高比例大于此值将被过滤。

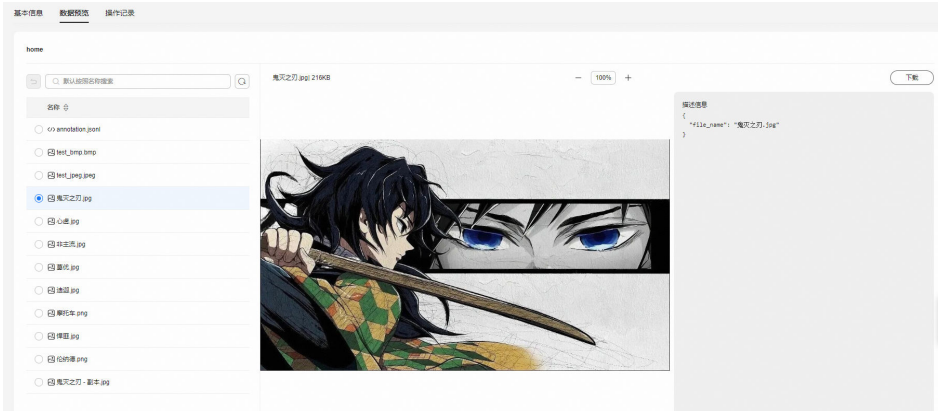
  最小文件大小：文件大小低于该文件大小会被过滤，单位为B。

- 参数配置样例：

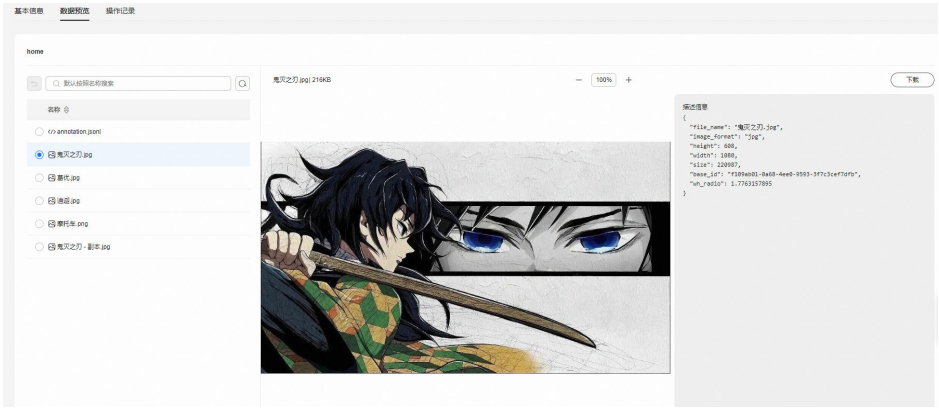
  

- 过滤样例：

  原数据集：

  过滤后：宽度低于1079的图片被过滤。



图片去重

- 适用的文件格式：
  jpg、jpeg、png、bmp。
  tar：所有图片保存为tar包。tar包含图片支持：jpg、jpeg、png、bmp图片类型。
- 各参数说明：
  待过滤内容类型：通过把图片结构化处理后，过滤重复的图片/图文对数据。
- 参数配置样例：
  不需要配置参数。
- 过滤样例：

图 4-40 精炼前

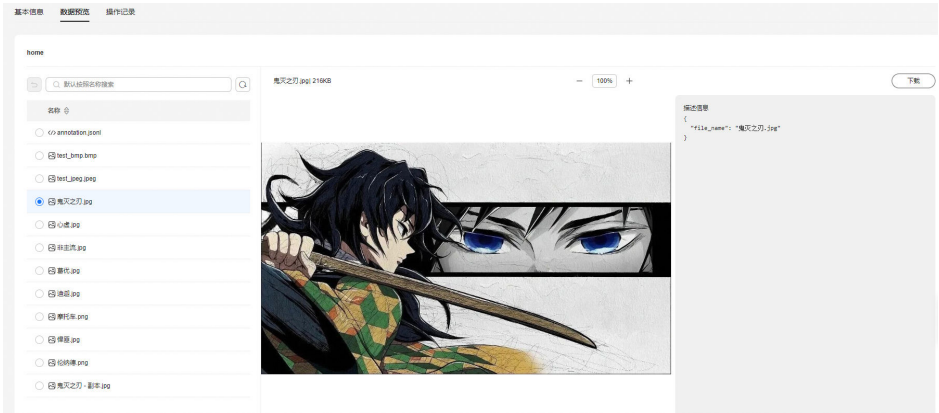
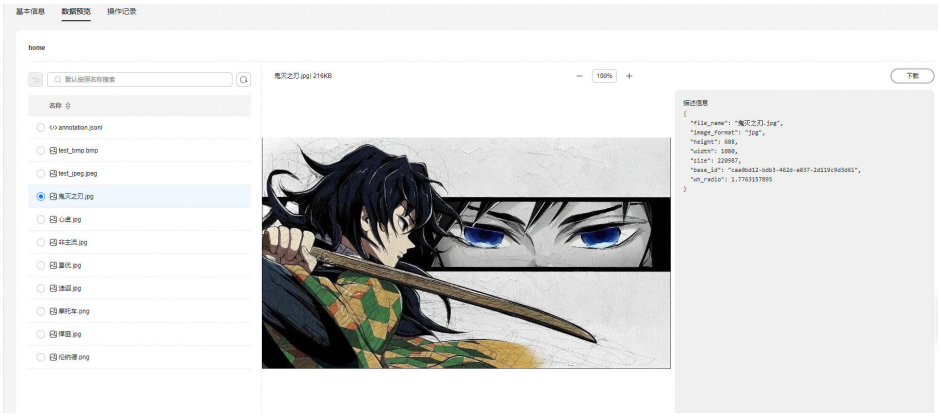


图 4-41 精炼后



色情图像检测

- 适用的文件格式：

  jpg、jpeg、png、bmp。

  tar：所有图片保存为tar包。tar包含图片支持：jpg、jpeg、png、bmp图片类型。

- 各参数说明：

  待打标内容类型：对图片的涉黄程度进行评分， 分数越高越危险。 评分范围(0.100)， 默认评分≥50分的视频可视为涉黄视频。

- 参数配置样例：

  是：开启过滤功能。

  否：关闭过滤功能。

- 检测样例：

  检测结果以image_porn对象存储在标注文件中。

  suggestion：对文件检测是否通过的结果，pass代表审核通过无相应的问题；review代表需要人工复核，您可以按照您的审核策略选择放通还是拦截；block代表待审文件存在问题。

  confidence：模型结果检测的置信度（注意这里的置信度代表模型给出建议的置信度）。如果suggestion为pass，则为零；如果suggestion为review/block，则为0-1。

  label：模型检测出的具体色情标签，如果未检测出则为空。

描述信息


```
{
  "base_id": "a93bc4e7-3166-4510-aa41-f603397c6690",
  "file_name": "色情图片.jpg",
  "height": 1279,
  "image_format": "jpg",
  "image_porn_label": {
    "suggestion": "review",
    "results": [
      {
        "category": "porn",
        "confidence": 0.5259,
        "label": "sexy"
      }
    ]
  },
  "size": 287324,
  "wh_radio": 0.712275215,
  "width": 911
}
```


危情图像检测

- 适用的文件格式：

  jpg、jpeg、png、bmp。

  tar：所有图片保存为tar包。tar包含图片支持：jpg、jpeg、png、bmp图片类型。

- 各参数说明：

  待打标内容类型：给危情图片内容打标签。

- 参数配置样例：

  不需要配置参数。

- 检测样例：检测结果以image_danger对象存储在标注文件中:

  suggestion：对文件检测是否通过的结果，pass代表审核通过无相应的问题；review代表需要人工复核，您可以按照您的审核策略选择放通还是拦截；block代表待审文件存在问题。

  confidence：模型结果检测的置信度（注意这里的置信度代表模型给出建议的置信度）。如果suggestion为pass，则为零；如果suggestion为review/block，则为0-1。

  label：模型检测出的具体危情标签，如果未检测出则为空。

描述信息


```
{
  "base_id": "86d6891d-732b-4d72-a2d6-ac20ec314ca8",
  "file_name": "持刀伤人v1.PNG",
  "height": 810,
  "image_danger_label": {
    "suggestion": "pass"
  },
  "image_format": "png",
  "image_terrorism_label": {
    "results": null,
    "suggestion": "pass"
  },
  "size": 923091,
  "wh_radio": 1.6259259259,
  "width": 1317
}
```


暴恐图像检测

- 适用的文件格式：

  jpg、jpeg、png、bmp。

  tar：所有图片保存为tar包。tar包含图片支持：jpg、jpeg、png、bmp图片类型。
- 各参数说明：

  待打标内容类型：过滤暴恐图像。
- 参数配置样例：

  是：开启过滤功能。

  否：关闭过滤功能。
- 使用场景：

  场景仅限暴恐相关场景，暂无法保证完全识别准确。
- 检测样例：检测结果以image_terrorism对象存储在标注文件中。

  suggestion：对文件检测是否通过的结果，pass代表审核通过无相应的问题；review代表需要人工复核，您可以按照您的审核策略选择放通还是拦截；block代表待审文件存在问题。

  confidence：模型结果检测的置信度（注意这里的置信度代表模型给出建议的置信度）。如果suggestion为pass，则为零；如果suggestion为review/block，则为0-1。

  label：模型检测出的具体暴恐标签，如果未检测出则为空。

描述信息


```
{
  "base_id": "8176bb22-d490-41f7-9ee3-2ca840669459",
  "file_name": "暴恐流血.jpg",
  "height": 500,
  "image_format": "jpg",
  "image_terrorism_label": {
    "suggestion": "block",
    "results": [
      {
        "category": "terrorism",
        "confidence": 0.998,
        "label": "bloody"
      }
    ]
  },
  "size": 86858,
  "wh_radio": 1.334,
  "width": 667
}
```


4.7 常见问题

1. **合成算子可以放在工作流中间吗?**

   不可以。当前版本合成算子仅支持放在工作流末端。如果需要对合成结果进行进一步处理，建议分两次任务执行：

   a. 第一次任务：加工算子 + 合成算子

   b. 第二次任务：对合成结果继续加工处理

2. **能否实现文本到图像的跨模态合成?**

   当前版本不支持跨模态合成。仅支持同模态的数据合成，例如：

   - 文本 → 文本（问答改写）
   - 文本 → 图像（不支持）
   - 图像 → 文本（不支持）

5 数据资产管理

5.1 数据资产介绍

数据资产是指在ModelArts平台中被纳入管理、存储并可供使用的数据集，包含**预置数据**和**我的数据**两类资产。

- **预置数据：**平台预置数据集是为用户提供的开箱即用的高质量数据资源，涵盖文本类、图片类、视频类和音频类四大类型的精选数据集。这些数据集经过严格筛选和预处理，可直接用于数据精炼、模型训练、微调和评估，大幅降低用户数据准备的时间成本和技术门槛。详见**预置数据**。
- **我的数据：**在控制台创建**数据连接**和**数据精炼**任务时，生成的数据集将作为数据资产放置在**我的数据**列表。详见**我的数据**。

5.2 预置数据

预置数据集是ModelArts平台为用户提供的开箱即用的高质量数据资源。这些数据集是行业通用数据集，符合开源协议要求，能够兼容主流的训练框架。使用预置数据集，能够保证数据集版本可追溯，确保实验可复现。您可以根据实际场景，选择合适的数据集直接在平台环境调用。

使用场景

预置数据典型使用场景如下：

1. 使用预置数据集配合用户自定义数据完成数据精炼，合成下游需要数据集。
2. 使用预置数据集完成大模型预训练与微调，提升模型基础能力，通过人类偏好数据优化模型响应质量。
3. 结合图像、视频、音频数据构建跨模态能力，开发多模态模型。
4. 作为标准测试集评估模型性能，完成模型能力基线评估。

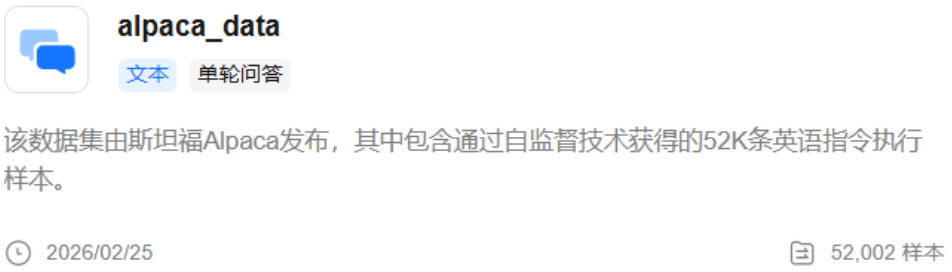
约束限制

- 仅西南-贵阳一区域的新版控制台支持。

操作指南

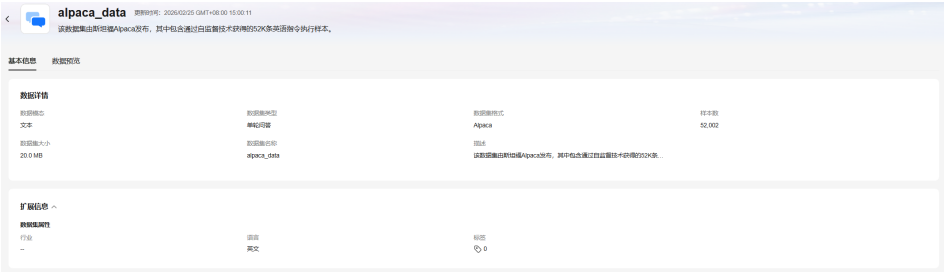
1. 前往**ModelArts管理控制台**。
2. 在左侧导航栏中选择“资产管理 > 数据 > 预置数据”页签，平台预置数据集会以卡片形式呈现。通过预置数据卡片，可查看**数据集名称、模态、类型、简介、更新时间、样本数**等信息。

图 5-1 预置数据集卡片



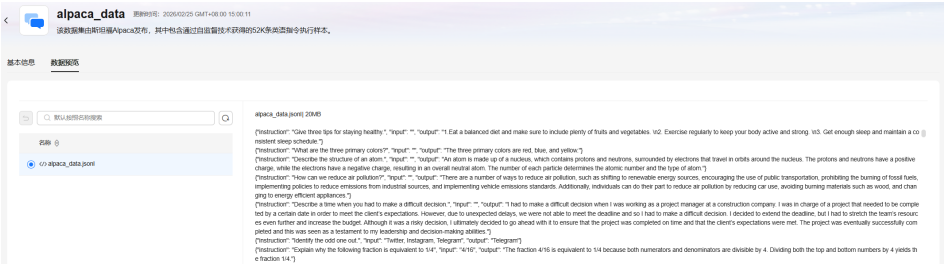
3. 单击预置数据集卡片，可查看预置数据集详情。包含**基本信息**和**数据预览**。
   - **基本信息：**预置数据集**名称、模态、类型、样本数、数据集大小、描述信息等**信息和**数据集属性、行业、语言、标签**等扩展信息。

图 5-2 预置数据基本信息



   - **数据预览：**数据预览能够支持文本、表格类结构化数据展示部分样例，支持分页查看、查看原始数据结构，非结构化数据（图像/音频）支持缩略预览。

图 5-3 预置数据预览



预置数据集介绍

ModelArts平台预置文本、图片类数据集，当前预置数据相关信息参见**表5-1**，请根据具体场景选择对应数据集。

表 5-1 预置数据集清单

名称	预置标签	数据集简介	大小	样本数	语言	链接
ai-expert-alpaca	文本、单轮问答	该数据集包含高质量的问答对，用于大型语言模型的监督式微调（SFT），重点关注三大核心人工智能技术领域：大型语言模型（LLM）、检索增强生成（RAG）和智能体系统。该数据集全面覆盖了这些前沿人工智能技术，涵盖英语和中文两种语言。	8.2 MB	11235	中文、英语	https://huggingface.co/datasets/GXMZU/ai-expert-alpaca?utm_source=chatgpt.com
GPT-4-LLM	文本、单轮问答	Alpaca-CoT是一个大规模、高质量、融合了多种任务类型的指令微调数据集。	33.47 MB	48818	中文	https://huggingface.co/datasets/QingyiSi/Alpaca-CoT/blob/main/alpacaGPT4/alpaca_gpt4_data.json
alpaca_data	文本、单轮问答	该数据集由斯坦福Alpaca发布，其中包含通过自监督技术获得的52K条英语指令执行样本。	20.0MB	52002	英语	https://huggingface.co/datasets/QingyiSi/Alpaca-CoT/tree/main/alpaca
alpaca_gpt4_data	文本、单轮问答	该数据集由Instruction-Tuning-with-GPT-4发布。它包含52K个由GPT-4生成的英语指令遵循样本，这些样本使用Alpaca提示词生成，用于微调大语言模型（LLMs）。	40.4MB	52002	英语	https://huggingface.co/datasets/QingyiSi/Alpaca-CoT/tree/main/alpacaGPT4
code_alpaca	文本、单轮问答	该数据集由codealpaca发布，包含20022个样本的代码生成任务。	6.7 MB	20022	英语	https://huggingface.co/datasets/QingyiSi/Alpaca-CoT/tree/main/CodeAlpaca

名称	预置标签	数据集简介	大小	样本数	语言	链接
lunara-aesthetic-image-variations	图片	该数据集包含Moonworks创作的原始图像和艺术作品。	17.7MB	36	中文	https://huggingface.co/datasets/moonworks/lunara-aesthetic-image-variations/tree/main

5.3 我的数据

在控制台创建**数据连接**和**数据精炼**任务时，生成的数据集将作为数据资产放置在**我的数据**列表。如果创建**数据连接**和**数据精炼**任务配置“立即上线数据集”选项，生成的数据集将自动上线为资产。如果没有勾选，则在资产清单是未上线的状态。

使用场景

我的数据典型使用场景如下：

1. 使用我的数据集完成数据精炼，生成下游需要数据集。
2. 使用我的数据集完成大模型预训练与微调，提升模型基础能力，通过人类偏好数据优化模型响应质量。
3. 作为我的数据集作为测试集评估模型性能，完成模型能力基线评估。

约束限制

- 仅西南-贵阳一区域的新版控制台支持。

操作指南

1. 前往ModelArts管理控制台，
2. 在左侧导航栏中选择“资产管理 > 数据 > 我的数据”，能够查看所有数据集和前用户自己创建的资产列表，也可以按照**数据集名称、数据模态、数据集类型、上线状态、创建者**维度过滤数据集资产。

图 5-4 过滤数据集资产

3. 单击搜索栏右侧“  ”图标，在右侧弹出的网页，可以设置搜索栏。可配置内容如表5-2所示。

表 5-2 我的数据清单配置

配置项	配置参数	配置说明
基础设置	表格内容折行	打开**自动折行**开关，单条数据资产项会扩行，数据信息能够完全展示。关闭开关，数据资产项不会扩行，数据信息可能显示不全。
	表格数据列固定	• **不固定**：数据资产记录如果超长支持拖动时，数据列均可拖动。 • **固定第一列**：数据资产记录如果超长支持拖动时，数据列第一列会冻结，其余数据列可拖动。 • **固定前两列**：数据资产记录如果超长支持拖动时，数据列第一列第二列会冻结，其余数据列可拖动。
	表格操作列固定	勾选**固定操作列**后，**操作**列固定在最后一列永久展示，不能调整操作列宽。
自定义显示列	设置展示清单选项	勾选需要展示的列名。**数据集名称**和**操作**是默认展示列，其余选项可以勾选是否展示。数据列名支持拖动调整顺序。

图 5-5 设置选项

4. 选择一个数据集，在“操作”列支持如下操作：
   - **上线。**未上线的数据集支持上线操作。单击“上线”，在弹出的对话框确认后，数据集完成上线。上线后的数据集能够作为后续开发的数据。
   - **下线。**已上线的数据集支持下线操作。单击“下线”，在弹出的对话框确认后，数据集完成下线。下线后的数据集不能作为后续开发的数据。
   - **删除。**数据集可被删除。删除后的数据集不是彻底删除，为避免误删，如果还想再继续使用，可以恢复数据集。对于已删除的数据，可以选择彻底删除，彻底删除后的数据集不可恢复。

图 5-6 已删除数据集

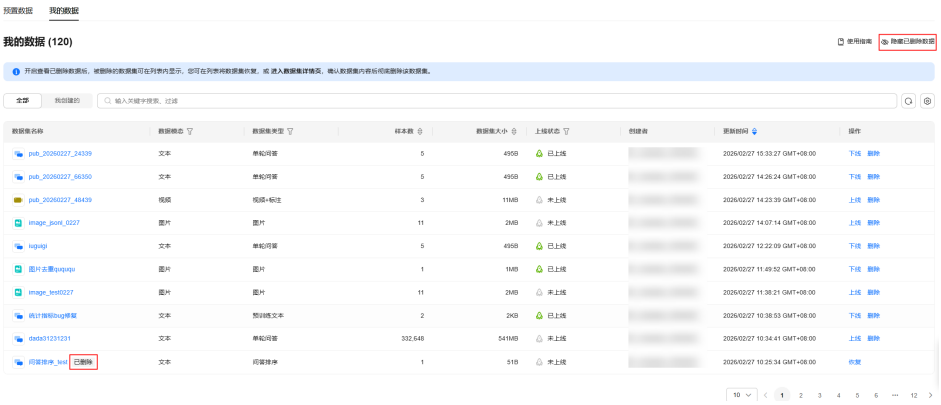
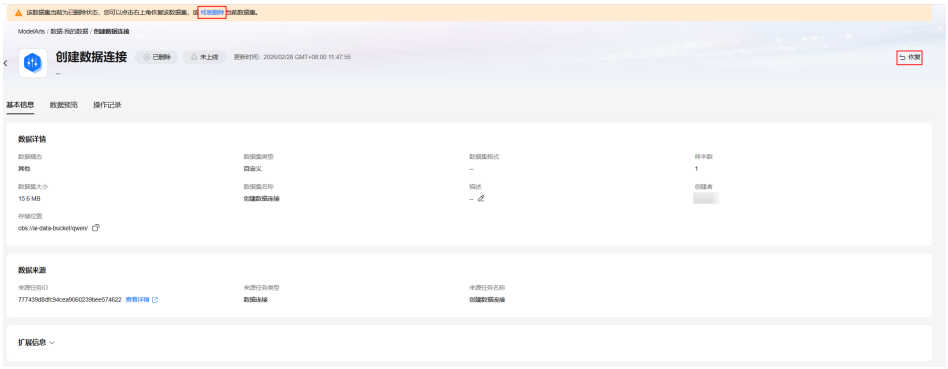


图 5-7 已删除数据集可恢复或彻底删除



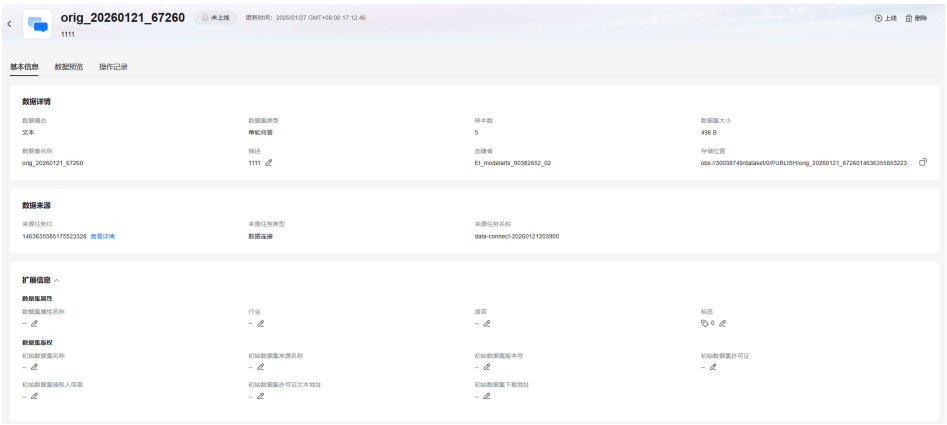
- **恢复。**对于已经删除的数据集，可以通过该选项恢复数据集。

数据资产详情管理

数据资产详情页面展示了当前数据集详细信息。在数据集工作区，单击任意数据集名称，就进入该数据集的详情页面。该页面有基本信息、数据预览和操作记录三个子页面，以下分别说明页面的作用和涉及的操作。在该页面右上角可以删除数据集。单击删除后，该数据集将彻底删除，请谨慎操作。

- **基本信息**。包含资产的信息如下：
  - **数据详情**：资产数据集**名称、模态、类型、样本数、数据集大小、描述信息、创建者、存储位置**等信息。
  - **数据来源**：数据资产生是什么任务生成，可以在**来源任务ID**链接到生成该数据的任务。
  - **扩展信息**：数据资产的属性及版权信息，支持手动修改该信息。

图 5-8 数据资产基本信息

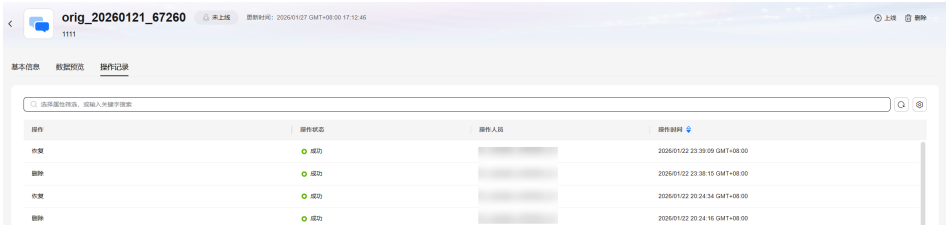


- **数据预览**。数据预览能够支持文本、表格类结构化数据展示3-5条典型样例，支持分页查看、查看原始数据结构，非结构化数据（图像/音频）支持缩略预览。支持数据集下载操作。

图 5-9 数据资产预览

- **操作记录。**在该界面可以查看“操作记录”。操作记录会记录当前数据集做过的所有操作。

图 5-10 数据资产操作记录



6 使用 CTS 审计 ModelArts 数据服务

ModelArts支持对接云审计CTS服务，通过云审计服务，您可以记录与ModelArts相关的操作事件，便于日后的查询、审计和回溯。

在您开启了云审计服务后，系统会记录ModelArts的相关操作，且控制台保存最近7天的操作记录。本节介绍如何在云审计服务管理控制台查看最近7天的操作记录。

前提条件

已开通云审计服务。详细操作请参见**《云审计服务用户指南》**。

数据准备支持审计的关键操作列表

数据准备支持审计关键操作如**表6-1**所示。

表 6-1 数据管理支持审计的关键操作列表

操作名称	资源类型	事件名称
创建数据集	dataset	createDataset
删除数据集	dataset	deleteDataset
更新数据集	dataset	updateDataset
发布数据集版本	dataset	publishDatasetVersion
删除数据集版本	dataset	deleteDatasetVersion
同步数据源	dataset	syncDataSource
导出数据集	dataset	exportDataFromDataset
创建数据集标签	dataset	createLabel
更新数据集标签	dataset	updateLabel
删除数据集标签	dataset	deleteLabel
删除数据集标签和对应的样本	dataset	deleteLabelWithSamples

操作名称	资源类型	事件名称
添加样本	dataset	uploadSamples
删除样本	dataset	deleteSamples

操作步骤

1. 登录云审计服务管理控制台。
2. 在管理控制台左上角单击 图标，选择区域。
3. 在左侧导航栏中，单击“事件列表”，进入“事件列表”页面。
4. 事件列表支持通过筛选来查询对应的操作事件。当前事件列表支持四个维度的组合查询，详细信息如下：
   - 事件来源、资源类型和筛选类型。
     在下拉框中选择查询条件。
     其中筛选类型选择事件名称时，还需选择某个具体的事件名称。
     选择资源ID时，还需输入某个具体的资源ID。
     选择资源名称时，还需选择或手动输入某个具体的资源名称。
   - 操作用户：在下拉框中选择某一具体的操作用户，此操作用户指用户级别，而非租户级别。
   - 事件级别：可选项为“所有事件级别”、“normal”、“warning”、“incident”，只可选择其中一项。
   - 时间范围：可选择查询最近七天内任意时间段的操作事件。
5. 在需要查看的事件左侧，单击 展开该事件的详细信息。
6. 单击需要查看的事件“操作”列的“查看事件”，可以在弹窗中查看该操作事件结构的详细信息。
   更多关于云审计服务事件结构的信息，请参见**《云审计服务用户指南》**。

