星海智算 · 开发文档

模型 API 接入手册

> 与在线站点同源生成的离线阅读版。实时模型规格、可调用权限和价格以平台当前返回为准。

来源版本 5236e2d4ce31e397
包含 59 个章节 · 在线入口 https://docs.jusuanhub.com
字体：HarmonyOS Sans，许可随 Markdown 包提供。

目录

开始调用模型
创建和管理 API Key
完成首个模型调用
API Key 异步媒体完整调用
发现模型与可用能力
文本对话与视觉理解
向量化与相关性重排
OCR 与文档解析
图片生成与参考图创作
Qwen Image 2.1
MiniMax H3 四种视频模式
H3 V2 直接调用
H3 · New API 中转站接入
Wan 角色动画与角色替换
选择其他视频能力
普通文生与图生视频
Bernini 参考视频生成
视频超分
X-Dub 人物口型同步
语音、音效与转写
提示词优化：H3 默认集成与 Qwen 手动改写
上传和读取素材
等待任务并下载结果
处理错误、限流和未知结果
预检媒体准入
上传输入素材
读取输入素材
读取生成结果
读取生成结果元信息
生成音频
转写音频
查询公共音色列表
查询公共音色详情
查询主账号积分余额
按变化游标查询调用明细

生成对话回复
将文档解析为结构化文件
生成文本向量
生成图片
查询异步任务
生成视频
优化媒体提示词
查询可用模型别名
识别文档文字
读取公开 API 协议
查询组织专属算力
查询组织与当前 Key 身份
同步组织门户用户映射
查询主账号配额与占用
查询组织可用服务状态
查询组织调用体验指标
重排候选文档
生成无状态 Responses 回复
查询服务与动态规格
H3 V2：查询视频任务
H3 V2：创建视频任务
H3 V2：读取生成视频
H3 V2：读取生成视频
模型 API 类型定义附录

/docs/models

开始调用模型

选择你的接入方式

外部程序使用 **API Key** 调用模型。托管在应用中心的应用应使用用户运行会话，不收集用户 Key，也不用开发者的 Key 为所有用户代付。

第一次接入，依次阅读 创建 API Key → 完成首个调用 → 处理失败和重试。

图片、视频和音频调用请直接使用 API Key 异步媒体完整调用：它明确了创建、鉴权查询、鉴权下载和 Web 后端转存的完整链路。

你想实现什么	调用指南
对话、视觉理解、图像续写	文本与视觉
搜索向量、相关性排序	向量与重排
提取文字、解析文档	OCR 与文档解析
生成图片、参考图创作	图片生成
生成视频、首尾帧、全能参考	MiniMax H3
角色动画与角色替换	Wan Animate
普通视频、参考视频、口型同步与超分	其他视频能力
语音、音效、转写	音频
在生成前优化提示词	提示词优化

接口参考提供请求字段、响应结构和错误定义；当前服务支持的合法规格以模型详情为准。

准备调用凭证

在平台的 API Key 页面创建凭证，并仅保存在服务端环境变量中。完整密钥仅在创建成功时显示一次，刷新或关闭后不能找回。

确认服务入口

从模型详情的 API 接入区域复制当前服务的公开 Base URL 和模型别名。`JUSP_BASE_URL` 包含末尾的 `/v1`，不要追加第二个 `/v1`。

```
export JUSP_BASE_URL='https://你的服务入口/v1'
export JUSP_API_KEY='<API_KEY>'
curl --fail-with-body "$JUSP_BASE_URL/models" \
  -H "Authorization: Bearer $JUSP_API_KEY"
```

模型发现请求用于检查授权范围，不会创建生成任务。不要将 Key 写入网页源码、公开仓库或截图。

生成并取得结果

文本调用可以同步返回。图片、视频等媒体任务需要先提交，再根据返回的任务地址查询终态，最后读取输出资源。请求被接受不代表生成已完成。

常见失败

现象	下一步
凭证被拒绝	检查密钥是否完整、有效，且覆盖当前服务。
请求受限	按接口返回的重试建议等待，不连续重复提交。
提交超时	先确认任务是否已经受理，不盲目新建生成任务。

`/docs/models/api-key`

创建和管理 API Key

在 Portal 创建

进入 模型广场 → API Key，点击“新建 API Key”。该入口需要登录，且账户满足平台当前准入要求。

1. 填写用途明确的名称，例如“服务端测试”。
2. 选择“平台内置服务集”或“单服务”。集合会随平台允许调用的服务变化；需固定最小授权时选择单服务。
3. 按需设置到期时间。
4. 创建成功后，在当前弹窗复制完整密钥并保存在服务端 secret 中，然后关闭弹窗。

完整 Key **只展示这一次**。列表中的前缀只用于识别，不能用于调用，也不能恢复完整 Key。不要截图、写入源码、浏览器存储或公开日志。

管理生命周期

列表支持分页和状态筛选，展示当前用户的密钥。冻结状态应联系平台支持，不能自行解冻。删除会永久停止后续调用，历史记录保留；不会承诺取消正在运行的任务。

密钥泄漏时立即删除，重新创建最小范围的 Key，再更新服务端配置。不要用重命名或隐藏前缀代替停用。

创建结果不确定

如果提交后连接中断，服务端可能已经创建成功。**不要连续重试创建**：先刷新列表查找刚才的名称和时间；即使存在，也不能从列表找回秘密。删除这次不可用的 Key 后再明确创建一次。

与应用 SDK 凭据区分

凭据	用途	不应当用于
模型 API Key	外部程序通过公开模型 API 调用	安装 SDK、替代应用用户会话
短期安装凭据	从受控索引下载 SDK / CLI	模型调用、应用运行
CLI 调试授权	为特定应用和候选版本测试真实能力	正式多用户运行
RuntimeSession	托管应用当前用户的运行身份	其他用户或其他应用的请求

下一步：完成首个模型调用。

/docs/models/quickstart

完成首个模型调用

1. 准备三个值

从 Portal 模型详情的“API 接入”取得 **Base URL** 和 **模型别名**，另在 API Key 页创建覆盖该模型的凭据。

`JUSP_BASE_URL` 包含末尾 `/v1`，没有末尾斜线。当前统一使用标准 HTTPS 443；不要沿用历史 `:10443` 地址。

```
export JUSP_BASE_URL='https://api.jusuanhub.com/v1'
export JUSP_API_KEY='<API_KEY>'
export JUSP_MODEL='<从模型详情复制的文本模型别名>'
```

这里的 Key 和模型值必须替换。避免 shell 调试回显，不在共享终端、录屏或日志中展示完整 Key。

2. 先检查可调用范围

```
curl --fail-with-body --max-time 30 "$JUSP_BASE_URL/models" \
  -H "Authorization: Bearer $JUSP_API_KEY"
```

这一步不会生成内容。失败时先处理鉴权或网络，不反复发起计费调用。服务发现见 GET /models 和 GET /services。

3. 发起文本请求

选择一个支持 Chat Completions 的模型。本例会调用真实模型，按当前服务规则计费；图片或视频模型请转到对应指南。

cURL

使用 `jq` 安全组装 JSON，不手动拼接用户文本。

```
jq -n --arg model "$JUSP_MODEL" \
  '{model:$model,messages:[{role:"user",content:"用一句话介绍成都"}]}' |
curl --fail-with-body --max-time 330 "$JUSP_BASE_URL/chat/completions" \
  -H "Authorization: Bearer $JUSP_API_KEY" \
  -H 'Content-Type: application/json' --data-binary @-
```

Python（标准库，无需平台 SDK）

```
import json
import os
import urllib.request

base = os.environ["JUSP_BASE_URL"].rstrip("/")
payload = {"model": os.environ["JUSP_MODEL"],
           "messages": [{"role": "user", "content": "用一句话介绍成都"}]}
request = urllib.request.Request(
    base + "/chat/completions", data=json.dumps(payload).encode(),
    headers={"Authorization": "Bearer " + os.environ["JUSP_API_KEY"],
             "Content-Type": "application/json"}, method="POST")
with urllib.request.urlopen(request, timeout=330) as response:
    result = json.load(response)
print(result["choices"][0]["message"]["content"])
```

JavaScript（服务端）

```
const base = process.env.JUSP_BASE_URL.replace(/\/$/, '');
const response = await fetch(`${base}/chat/completions`, {
  method: 'POST', signal: AbortSignal.timeout(330_000),
  headers: { Authorization: `Bearer ${process.env.JUSP_API_KEY}`, 'Content-Type': 'application/json' },
  body: JSON.stringify({ model: process.env.JUSP_MODEL,
    messages: [{ role: 'user', content: '用一句话介绍成都' }] }),
});
if (!response.ok) throw new Error(`调用失败：HTTP ${response.status}`);
const result = await response.json();
console.log(result.choices[0].message.content);
```

这些超时值只是客户端等待预算，不扩大模型输出能力，也不会使服务端工作自动取消。模型输出上限、流式能力和更长调用期限应以目标服务公开合同为准。

本页 Python 示例只使用标准库。模型详情生成的媒体流程 Python 示例使用 `requests`，运行前在自己的虚拟环境安装该依赖；cURL 流程使用 `jq`，JavaScript 流程在支持内置 `fetch` 的 Node.js 服务端运行。它们都不是需要私有安装凭据的平台应用 SDK。

图片、视频和音频的**生成类服务**使用异步任务：创建成功后还必须鉴权查询 `jobUrl`、再鉴权下载 `contentUrl`。可直接复制 API Key 异步媒体完整调用 中的 cURL 或 Java 服务端示例。图片和视频的**理解类服务**（例如 Qwen3.8 识图）使用 `/chat/completions`，返回 JSON 或 SSE；图片输入与 WorkBuddy 配置见文本与视觉。

4. 判断是否完整

读取 `choices[].finish_reason`。长度限制结束不等于内容已经完整；检查目标模型允许的 `max_tokens`，必要时分段生成。流式调用还须正确消费 SSE，详见 文本与视觉。

请求失败或超时不能无条件重放，尤其已经开始返回内容时。参见 错误与恢复。

`/docs/models/api-key-media`

API Key 异步媒体完整调用

Account API Key 调用 H3 时，**生成任务默认包含提示词优化**：创建一次任务，平台先优化、再生成视频。无需额外调用优化接口；优化按实际用量单独计费。

要直接使用提交的文本，在生成请求顶层传 `"optimizePrompt": false`。它只接受 JSON 布尔值，不能写成字符串 `"false"`。关闭优化仍需内容安全审核；如果已单独优化并采用文本，也应传 `false`，避免重复优化。独立调用见提示词优化。

创建、查询和下载是同一条受保护调用链，后两步仍要携带创建任务时使用的 Account API Key。

调用规则

规则	正确做法
提示词优化	H3 缺省或 `true` 开启；顶层 `optimizePrompt: false` 关闭。默认流程只创建一个 Job。
API 入口	固定为 `https://api.jusuanhub.com/v1`，不使用 Portal，也不带历史 `:10443`。
任务地址	保存并原样使用创建响应中的绝对 `jobUrl`；不要删除 `model` 查询参数。
结果地址	保存并原样使用终态响应中的 `outputs[].contentUrl`。
鉴权	创建、查询和下载均发送 `Authorization: Bearer $JUSP_API_KEY`。
水印	API Key 原样下载默认返回无可见水印的 source；只有显式 `deliveryMode=marked` 才生成带平台标识版本。

`contentUrl` 不是匿名公开 URL。未带 Key 的请求会以隐私保护方式返回 404，避免向未知调用方泄露资产是否存在。这个 404 不足以证明任务或资产不存在。

资源保存与空间配额

普通 Account API Key 与门户体验中心使用同一个人资源空间。主动上传和生成文件默认保存 **30 天**，分别从上传成功入库、产物成功提交起算；查询、重复上传、幂等重试和再次引用不会重新计时。在门户“我的资源”中可查看入库、到期时间并显式永久保存，永久保存仍占用空间。历史未记录新保存策略的资源在管理员明确生效前保持原策略；生效后从实际转换时刻起算，保留原入库时间及永久保存状态，不根据旧创建时间追溯到期。

选择	请求方式	保存与展示
默认保存	省略 `storageMode`，或传 `personal_30d`	进入个人资源空间，首次入库起 30 天；可在门户永久保存。
一小时临时保存	显式传 `temporary_1h`	需要平台已允许该账号或所属组织使用；不出现在个人长期资源列表，仍计入实际空间占用。
公网 URL 参考输入	使用支持的 URL 输入协议	参考输入遵循独立的一小时任务临时合同，不自动进入个人资源库；生成结果仍按请求的保存方式处理。

`storageMode` 是平台保存选项，支持图片、音频、视频及文档解析的 JSON 生成/处理请求，也支持主动 multipart 上传的同名表单字段；MiniMax V2 的输出支持同一选项。上传和生成是两次独立请求，分别选择保存方式。未获允许或策略暂不可确认时，临时模式请求会被拒绝；已接受任务保留受理时的选择。输入、输出分别从各自成功入库起保留一小时。到期停止业务读取，物理文件由后台异步回收，执行中的必要引用保护仍生效。

资产响应中的 `storageMode`、`ingestedAt` 和 `expiresAt` 用于核对保存方式与时间；已在门户永久保存的资源没有自动到期时间。不要把旧响应中的到期时间当作永久保存后的当前状态。

空间满额时，新上传和新生成会被拒绝；已有资源仍可查看、下载、删除，原任务重放和已接受生成的交付继续完成。回收站、临时资源及尚待物理回收的文件仍占空间，后台诊断副本不重复计入个人用量。普通 Key 不能通过应用来源或隐藏产物字段绕过保存与配额。

1. 创建 H3 任务

先在服务端设置环境变量。Key 不得进入网页源码、浏览器存储、URL 或日志。

```
export JUSP_BASE_URL='https://api.jusuanhub.com/v1'
export JUSP_API_KEY='<API_KEY>'
export JUSP_MODEL='minimax-h3'
export JUSP_IDEMPOTENCY_KEY="$(openssl rand -hex 16)"
```

文生视频最小请求：

```
jq -n --arg model "$JUSP_MODEL" '{
  model:$model,
  generationMode:"t2v",
  prompt:"清晨的海岸，镜头缓慢推进，柔和的自然光",
  resolutionTier:"768p",
  orientation:"landscape",
  aspectRatio:"16:9",
  seconds:5
}' |
curl --fail-with-body --max-time 60 "$JUSP_BASE_URL/media/generations" \
  -H "Authorization: Bearer $JUSP_API_KEY" \
  -H "Idempotency-Key: $JUSP_IDEMPOTENCY_KEY" \
  -H 'Content-Type: application/json' \
  --data-binary @- --output create.json
```

H3 四种模式缺省开启 `optimizePrompt`，先完成提示词准备再生成。已独立优化并采用最终文本，或希望直接使用原文时，在创建正文显式添加 `"optimizePrompt": false`，优化调用和费用为零。该开关只接受布尔值；省略不会让其他模型自动优化。重试与压缩按实际优化用量计费，完整上界和人工采用例子见提示词优化。

平台尚未启用自动优化能力时，H3 缺省或 `optimizePrompt: true` 请求返回 `503 prompt_optimization_capability_unavailable`。此时不会按原文自动生成；需要直接使用已确认文本时显式传 `false`，仍需满足正常生成审核与规格要求。默认优化示例以平台已启用该能力为前提。

创建成功通常返回 202。保存整个响应，尤其是绝对 `jobUrl`：

```
{
  "jobId": "job_...",
  "jobUrl": "https://api.jusuanhub.com/v1/jobs/job_...?model=minimax-h3",
  "status": "queued"
}
```

2. 查询返回的 jobUrl

不要把 `jobId` 拼到 Portal 域名。优先读取响应中的 URL:

```
JOB_URL="$(jq -er '.jobUrl' create.json)"
curl --fail-with-body --max-time 60 "$JOB_URL" \
  -H "Authorization: Bearer $JUSP_API_KEY" \
  --output job.json
```

202 只代表受理，包含自动优化准备的时间。若优化技术重试一次后仍失败，查询 Job 得到 `failed` 和 `errorCode: prompt_optimization_failed`，不会回退原文继续生成；查看费用回执，不能把失败等同未产生优化费用。

`queued` 或 `running` 时按 `retryAfterSeconds` / `refreshAfterSeconds` 有界轮询；`succeeded`、`failed`、`canceled` 是终态。本地等待超时不会取消服务端任务，更不能换新幂等键重新创建。

自动优化的错误会区分原因：准备超时为 `prompt_optimization_timeout`；内容审核未通过为 `content_safety_blocked`；需要补充说明为 `content_safety_requires_clarification`；优化稿被拒绝为 `prompt_optimization_candidate_rejected`；审核服务暂不可用为 `content_safety_unavailable`。这些错误发生于 `failureStage: prompt_optimization`，安全相关原因的 `phase` 为 `content_safety`。技术失败继续使用 `prompt_optimization_failed`。响应说明视频生成尚未开始，优化费用仍按实际调用结算。

查询 HTTP 200 仅表示查询成功，必须检查正文 `status`、`errorCode`、`errorMessage`；有 `details` 时可显示受控的审核原因和操作指导。`retryable: false` 表示不要自动新建付费任务。恢复查询保留原失败原因。

3. 鉴权下载 contentUrl

任务成功后，直接读取返回的资源地址。省略 `deliveryMode` 默认返回无可见水印的 source master，通常在 Job `succeeded` 后立即可下载:

```
CONTENT_URL="$(jq -er '.outputs[0].contentUrl // .assets[0].contentUrl' job.json)"
curl --fail-with-body --location --max-time 300 "$CONTENT_URL" \
  -H "Authorization: Bearer $JUSP_API_KEY" \
  --output result.mp4
```

需要带平台标识版本时才显式追加 `deliveryMode=marked`。该版本由独立导出任务生成，可能先返回 HTTP 202；不能把 202 JSON 保存成 MP4:

```
MARKED_URL="${CONTENT_URL}&deliveryMode=marked"
while :; do
  STATUS="$(curl --silent --show-error --location --max-time 300 \
    -H "Authorization: Bearer $JUSP_API_KEY" \
    --dump-header marked.headers \
    --output marked.body \
    --write-out '%{http_code}' \
    "$MARKED_URL")"
  if [ "$STATUS" = 200 ]; then
    mv marked.body result-marked.mp4
    break
  fi
  if [ "$STATUS" != 202 ]; then
    printf '带标识下载失败，HTTP %s\n' "$STATUS" >&2
    cat marked.body >&2
    exit 1
  fi
  WAIT_SECONDS="$(jq -r '.retryAfterSeconds // 2' marked.body)"
  sleep "$WAIT_SECONDS"
done
```

`deliveryMode` 只接受 `source` 和 `marked`；省略等同 `source`。它是结果交付选项，不属于 H3 `mediaSpec`，不会重新调用模型或重复计费。

不要把 `contentUrl` 直接赋给 `<video src>` 或 `<img src>`：浏览器无法安全持有 Account API Key。Web 应用应由自己的后端下载/流式转发，或转存到应用对象存储后返回应用自己的短期签名 URL。

Java / Spring 服务端

下面使用 JDK `HttpClient` 和 Jackson，可直接放进 Spring Service。重点是查询和下载继续复用同一个 Authorization 头，并只向可信 API 来源发送凭据。

```
import com.fasterxml.jackson.databind.JsonNode;
import com.fasterxml.jackson.databind.ObjectMapper;
import java.net.URI;
import java.net.http.HttpClient;
import java.net.http.HttpRequest;
import java.net.http.HttpResponse;
import java.nio.file.Path;
import java.time.Duration;
import java.util.Set;
import java.util.UUID;

public final class H3Example {
  private static final URI BASE = URI.create("https://api.jusuanhub.com/v1");
  private final String key = System.getenv("JUSP_API_KEY");
  private final ObjectMapper json = new ObjectMapper();
  private final HttpClient http = HttpClient.newBuilder()
      .connectTimeout(Duration.ofSeconds(10)).build();

  private HttpRequest.Builder authenticated(URI uri) {
    if (!uri.getScheme().equals(BASE.getScheme()) ||
        !uri.getAuthority().equals(BASE.getAuthority()))
```

```
      throw new IllegalArgumentException("拒绝向非 API 来源发送 Key");
    return HttpRequest.newBuilder(uri)
        .timeout(Duration.ofSeconds(60))
        .header("Authorization", "Bearer " + key);
  }

  public void run() throws Exception {
    var payload = json.createObjectNode()
        .put("model", "minimax-h3")
        .put("generationMode", "t2v")
        .put("prompt", "清晨的海岸，镜头缓慢推进，柔和的自然光")
        .put("resolutionTier", "768p")
        .put("orientation", "landscape")
        .put("aspectRatio", "16:9")
        .put("seconds", 5);
    var createRequest = authenticated(BASE.resolve("/v1/media/generations"))
        .header("Content-Type", "application/json")
        .header("Idempotency-Key", UUID.randomUUID().toString())
        .POST(HttpRequest.BodyPublishers.ofString(json.writeValueAsString(payload)))
        .build();
    var createResponse = http.send(createRequest, HttpResponse.BodyHandlers.ofString());
    if (createResponse.statusCode() / 100 != 2)
      throw new IllegalStateException(createResponse.body());
    JsonNode job = json.readTree(createResponse.body());
    URI jobUrl = URI.create(job.path("jobUrl").asText());

    var terminal = Set.of("succeeded", "failed", "canceled");
    var deadline = System.nanoTime() + Duration.ofMinutes(30).toNanos();
    while (!terminal.contains(job.path("status").asText())) {
      if (System.nanoTime() >= deadline)
        throw new IllegalStateException("本地等待结束；保留 jobUrl 稍后恢复");
      Thread.sleep(Math.max(1, job.path("retryAfterSeconds").asInt(3)) * 1000L);
      var response = http.send(
          authenticated(jobUrl).GET().build(), HttpResponse.BodyHandlers.ofString());
      if (response.statusCode() / 100 != 2)
        throw new IllegalStateException(response.body());
      job = json.readTree(response.body());
    }
    if (!job.path("status").asText().equals("succeeded"))
      throw new IllegalStateException(job.toString());
    // 原样 contentUrl 默认返回无可见水印的 source master。
    URI contentUrl = URI.create(job.path("outputs").get(0).path("contentUrl").asText());
    var download = http.send(
        authenticated(contentUrl).timeout(Duration.ofMinutes(5)).GET().build(),
        HttpResponse.BodyHandlers.ofFile(Path.of("result.mp4")));
    if (download.statusCode() / 100 != 2)
      throw new IllegalStateException("下载失败：" + download.statusCode());
  }
}
```

正式项目应把 `HttpClient`、ObjectMapper 和 API 配置作为单例依赖注入，并按 错误与恢复 处理 429、5xx、超时和幂等恢复。

快速排错

现象	根因和处理
`api.jusuanhub.com:10443` 失败	删除旧端口，改用标准 443；同时检查鉴权和查询参数。
`portal.jusuanhub.com/v1/...` 返回 HTML 403/405	Portal 是网页入口，不是模型 API；改用 `api.jusuanhub.com/v1`。
Job 或资源返回 404，但 Portal 中能看到结果	通常缺少 Bearer、Key 不属于原任务，或手工拼 URL 丢了 `model`；原样使用返回 URL。
下载得到很小的 JSON 而不是视频	只有显式 `deliveryMode=marked` 才可能返回 202；读取 `retryAfterSeconds` 后重试同一 URL，不能把 202 body 当作媒体文件。
需要带平台标识结果	在原 `contentUrl` 后追加 `&deliveryMode=marked`；默认不带可见水印。
409	相同幂等键对应的状态/正文冲突；按原键恢复，开始新业务动作才换键。
429	按 `Retry-After` 退避，不并发重放。
5xx 或网络超时	写请求结果可能未知；先用原 `jobUrl` / 幂等键查询，不能立即创建第二个任务。

H3 四种模式和素材限制见 MiniMax H3，通用任务状态机见 等待任务并下载结果。

/docs/models/services

发现模型与可用能力

先设置 快速开始 中的环境变量。服务发现是只读操作，不生成内容；它确认的是当前身份能看到的服务，不保证稍后的生成一定通过准入。

查询服务

```
curl --fail-with-body --max-time 30 "$JUSP_BASE_URL/services" \
  -H "Authorization: Bearer $JUSP_API_KEY" --output services.json
jq '.data[] | {modelAlias, serviceDisplayName, serviceKind, available, endpoints, mediaSpecContract, apiUsageContract,
promptOptimization}' services.json
```

字段	怎么使用
`modelAlias`	后续请求的 `model`，不是内部服务 ID
`serviceDisplayName`	展示给用户的服务名称，不替代别名
`serviceKind`	区分文本、图片、视频、音频等能力
`available`	当前可调用性；字段缺失不等于可用，提交前仍会进行实际准入
`endpoints`	当前服务公开支持的操作入口
`mediaSpecContract`	当前生成模式、输入素材、合法规格与大小限制
`apiUsageContract`	当前服务的参数说明和调用示例
`promptOptimization`	是否开放优化、支持模式及参考素材约束；缺失时不自行假设支持

简化的模型目录见 GET /models；精确字段见 GET /services。保留响应中不认识的扩展字段，不据此拼装内部地址。

如何选择参数

1. 在返回集合中找到目标公开别名，核对能力类别与可调用性。
2. 从该服务当前合同选择模式，再按模式要求上传素材。
3. 仅选该模式接受的输入、清晰度、画幅、数量或时长；不生成所有选项的笛卡尔积。
4. 使用 `apiUsageContract` 与对应教程形成请求；公共 OpenAPI 的字段全集不表示每个模型都接受全部字段。

例如 Wan 角色动画不需要提示词，视频超分不接受任意时长，H3 四模式的素材要求不同。见 选择视频能力。

目录变化与失败恢复

服务会下架、更新规格或变为暂不可用；可以短期缓存发现结果，但不能将它当作永久授权。收到规格不匹配时重新发现并让用户确认新选择，不能自动切换模型或提高付费规格。401/403 先处理凭据或范围；空列表先核对 Key 的授权服务，不据此断言平台没有模型。

文本对话与视觉理解

选择模型和消息

文本与支持视觉的模型都使用 POST /chat/completions。model 是公开模型别名，messages 是消息数组。先完成 快速开始，再增加目标模型支持的参数。

模型的上下文、单次输出上限和默认值不是同一个概念。不要把上下文长度作为 max_tokens；不要将某个模型的默认输出上限推广到所有服务。模型详情提供服务当前开放的参数范围。

带图片的消息

先按 素材上传 上传图片，取回 asset.assetId。支持视觉输入的服务使用以下消息内容，素材必须属于当前调用身份并适配同一模型：

```
{
  "model": "<MODEL_ALIAS>",
  "messages": [{
    "role": "user",
    "content": [
      {"type": "text", "text": "按画面顺序描述这个场景"},
      {"type": "image_url", "image_url": {"url": "asset://<ASSET_ID>"}}
    ]
  }]
}
```

只有服务明确声明时才可使用外部图片 URL 或内联 Base64；Qwen3.8 的兼容格式见下文。其他服务继续使用素材上传与 asset:// 引用。图像续写也由同一真实模型能力决定；体验中心中的模板是提示词辅助，不是特殊 API。

消费流式响应

只有目标服务声明支持流式时才设置 stream: true。cURL 加 --no-buffer；服务端客户端按 SSE 事件边界解析，不按 TCP 数据块直接 JSON 解码。

1. 确認 HTTP 成功，响应类型为 text/event-stream。
2. 累积数据直到完整事件边界，再解析 data:；一个网络块可能包含半条或多条事件。
3. 合并 choices[].delta.content，保留 finish_reason；Qwen3.8 开启思考时另行累计 delta.reasoning_content，不要混入最终答案。
4. 只有收到 data: [DONE] 才认为流传输正常结束；若 finish_reason=length，仍需向用户说明内容受到长度限制。
5. 连接提前关闭时保留“部分内容”状态，不标记成功，不自动把同一请求再发一次。

服务端期限与客户端等待预算分别生效。HTTP 响应开始前的超时可能是 504；已开始的 SSE 出错可能以连接关闭表现，不能只依赖 HTTP 状态。

可运行的流式示例

下载服务端 JavaScript 示例，设置快速开始的三个环境变量后运行 `node chat-stream.mjs`。需要支持内置 `fetch` 的 Node.js，不能在浏览器中运行或把 Key 编译到网页里。

示例处理 UTF-8 跨网络块、LF/CRLF 事件行、连续事件、`[DONE]`、流内错误和非正常关闭；`finish_reason=length` 会明确提醒内容受到长度限制，不将它显示为完整回答。示例不会自动重试计费调用。它在普通模式使用 330 秒、思考模式使用 1900 秒客户端等待预算，真实服务允许的时限和输出范围仍以模型合同为准。

请求字段和响应 · 错误恢复。

进阶多模态内容分析

`qwen3-6-35b-a3b` 使用原“进阶多模态内容分析”服务入口。模型广场的服务卡片在 **API 接入** 中提供当前合同与 cURL、Python、Node.js 示例；旧 Console 和新 Portal 使用相同的公开模型别名。以下能力仅适用于服务当前声明的合同：

项目	公开范围
素材组合	纯文本、最多 5 张图片、最多 2 个视频，或最多 2 张图片与 2 个视频混合
素材大小	单图 16MiB、单视频 128MiB；每视频最多采样 16 帧
Token 预算	输入最多 65536，输出最多 8192，合计最多 73728；`max_tokens` 是输出预算，模型可能提前结束
响应	非流式 JSON 或 `stream: true` 的 SSE
函数工具	`tools`、`tool_choice` 和 `parallel_tool_calls`；可与视觉输入组合
自定义停止	`stop` 为字符串或最多四项的字符串数组；匹配的停止文本不回传

本节 `qwen3-6-35b-a3b` 不接受调用方扩大素材预算、开启思考模式或传入内部结构化生成控制项；Qwen3.8 的思考用法见下文。提示词优化器使用的模型由平台独立配置，此服务升级不意味着优化器已切换。

视频消息把图片示例的内容项换成 `{"type":"video_url","video_url":{"url":"asset://<VIDEO_ASSET_ID>"}}`。多素材按 `content` 数组顺序组合；每个素材都先以同一身份、同一模型上传。超出预算会拒绝，不能依赖服务自动丢图或缩减合法引用。

“多模态内容分析”和“进阶多模态内容分析”启用扩展版本后，可在一次消息中提交9张纯图片或8张图片与1个视频。旧版本仍按服务卡片声明的上限接收。 扩展版本保留2个视频及2图 + 2视频的用法，总媒体最多9项；包含2个视频时最多配2张图。数量按全部消息累计，单素材大小、采样帧数和 token 预算同时生效。 每张图片使用独立 `image_url`，视频使用独立 `video_url`，无需拼图；SDK 同步与异步调用使用同一消息结构。

Qwen3.8 多模态理解

公开模型别名为 `qwen38-27b`，使用同一个 `/v1/chat/completions` 接口。调用前在模型广场确认服务可用及当前合同。图片和视频混合时，数量上限随视频个数变化。

项目	当前范围
Token 预算	输入最多 262080，输出上限为 262143，输入输出合计最多 262144（256K）；默认输出预算 64
素材	纯图片最多 9 张；1 个视频可搭配最多 8 张图片，2 个视频可搭配最多 2 张图片；总媒体不超过 9 项，按整次请求累计
大小与格式	单图 16MiB（JPEG、PNG、WebP），单视频 128MiB（MP4），每视频最多采样 16 帧
视觉处理	单张图片或视频帧解码后最多 16777216 像素；最长边缩放至不超过 1024，再由视觉处理器处理；每帧最多 4096 个 patch，总视觉 tokens 最多 16384
素材合计预算	媒体文件总量最多 384MiB；原图和采样视频帧累计最多 134217728 像素；预处理按“媒体字节数 + 缩放后总像素数×24”计算，最多 512MiB
输出	非流式 JSON、SSE；可设置 `stream_options.include_usage: true` 获取流式用量
思考模式	默认关闭；支持 `enable_thinking`、`reasoning_effort`，思考与最终答案分别返回，见思考模式
函数工具	支持纯文本、图片和视频历史与工具组合，支持 JSON/SSE；以目标服务公开的 `visionChatContract.toolsTextOnly` 为准
自定义停止	`stop` 仅支持纯文本非流式请求
并发	每个四卡后端最多 8 个活动请求；账号和空间额度仍独立生效，长请求还受共享容量调度
结构化输出	`json_schema`：1 ~ 16 个必填字符串字段，禁止额外字段；非流式、关闭 thinking，不与工具或 stop 组合
历史缓存	同一可信调用者可复用已完成历史；后端本地缓存可能未命中，实际命中见 `usage.prompt_tokens_details.cached_tokens`

输入和输出上限不能分别取满：例如输入 262080 tokens 时，输出预算最多剩余 64。显式设置适合业务的 `max_tokens`；默认 64 很容易让较长回答以 `finish_reason=length` 结束。素材同时消耗上下文预算，文件大小合规不表示一定满足 token 预算。不要把历史测试图片或视频的尺寸当作任意尺寸都可接收的保证；不要传入私有预处理参数扩大服务合同。

```
curl --no-buffer --max-time 1900 https://api.jusuanhub.com/v1/chat/completions \
  -H "Authorization: Bearer $JUSP_API_KEY" \
  -H 'Content-Type: application/json' \
  -d '{"model":"qwen38-27b","messages":[{"role":"user","content":"用三句话解释推理服务的并发限
制。"}],"max_tokens":256,"stream":true,"stream_options":{"include_usage":true}}'
```

思考模式

在普通 Qwen3.8 对话中设置 `enable_thinking: true`；也可只设置非 `none` 的 `reasoning_effort` 来开启。两者都省略时关闭。它们是请求 JSON 的顶层字段，不放在 `messages`、提示词文字或自定义请求头中；其他模型不能据此推定支持思考。

参数与默认值	用法
`enable_thinking`（boolean，默认 `false`）	`true` 开启，`false` 关闭；不要使用字符串 `"true"`
`reasoning_effort`（string，可省略）	`none` 关闭；`minimal` / `low` 映射为 low，`medium` 为 medium，`high` / `xhigh` / `max` 映射为 xhigh；仅设置 `enable_thinking: true` 时采用 xhigh
`preserve_thinking`（boolean，默认 `true`）	控制聊天模板对已有思考历史的保留，不提供服务端会话或缓存保证
`chat_template_kwargs`（object，可省略）	兼容嵌套传入以上三个字段；优先使用顶层写法，不要在两个位置重复设置

同时设置开关和档位时必须一致：`true` 可配非 `none` 档位，`false` 可配 `none`；冲突返回 400。顶层与 `chat_template_kwargs` 的同名字段也不能冲突。上述别名实际归为三个思考档位，不表示七种独立计算预算。

下面是非流式调用。将 `JUSP_API_KEY` 设置为自己的 Key；`max_tokens` 同时覆盖思考和最终回答，不能再沿用默认的 64 tokens 期待完整推理回答。

```
curl --max-time 1900 https://api.jusuanhub.com/v1/chat/completions \
  -H "Authorization: Bearer $JUSP_API_KEY" \
  -H 'Content-Type: application/json' \
  -d '{"model":"qwen38-27b","messages":[{"role":"user","content":"比较两个方案，并给出推荐理
由。"}],"enable_thinking":true,"reasoning_effort":"medium","max_tokens":4096,"stream":false}'
```

非流式从 `choices[0].message.reasoning_content` 读取思考，从 `choices[0].message.content` 读取最终答案；两者分别展示，不拼接为答案。关闭思考时思考字段可能不存在。流式对应 `choices[].delta.reasoning_content` 和 `choices[].delta.content`，两类增量应分别累计；只读 `delta.content` 的客户端在思考阶段可能暂时没有可展示的答案。

将请求改为 `stream: true` 并加入 `stream_options: {"include_usage": true}` 即可使用 SSE。使用本页服务端 JavaScript 示例时设置 `JUSP_MODEL=qwen38-27b`、`JUSP_ENABLE_THINKING=true`，可选 `JUSP_REASONING_EFFORT=medium`；已有的 `JUSP_BASE_URL` 和 Key 设置沿用快速开始。示例解析器提供两类增量回调并保留最终 usage；命令行只输出答案，思考保留在返回值的 `reasoning` 中，不在文档网页执行计费调用。

`finish_reason=length` 表示输出预算耗尽，可能只有思考而没有最终答案；即使收到 `[DONE]` 也不能标记为完整回答。思考消耗包含在 `usage.completion_tokens` 中，不另加一次；用量事件可能没有 choices，须继续读到结束。HTTP 200 不代替流完成，连接中断不自动重放。开启思考可能增加等待和输出用量，`max_tokens` 仍受输入输出总预算约束。

思考可用于普通文本与所选服务支持的视觉、工具对话；工具结果按下文两跳协议提交。`response_format.type=json_schema` 必须关闭思考与流式，不组合 tools 或 stop。关闭时省略档位，或同时使用 `enable_thinking: false` 与 `reasoning_effort: "none"`。多轮调用保留原始 assistant 消息及其中已有的 `reasoning_content`、`tool_calls`，不要把思考正文改写成 user 消息。

结构化输出与历史缓存

结构化调用示例：

```
{"model":"qwen38-27b","messages":[{"role":"user","content":"将春季新品发布计划整理为主题与摘
要。"}],"enable_thinking":false,"stream":false,"max_tokens":512,"response_format":{"type":"json_schema","json_schema":
{"name":"summary","strict":true,"schema":{"type":"object","properties":{"topic":{"type":"string"},"summary":
{"type":"string"}},"required":["topic","summary"],"additionalProperties":false}}}}
```

历史缓存不需要提交 session ID 或缓存身份请求头。继续对话时保留完整、有序的历史消息和模型原始回答，再追加新问题。只有精确历史匹配且命中同一后端的有效缓存才可能复用；不同 Key、调用者、应用和工作空间隔离。`cached_tokens` 是输入 tokens 的子集，不另外加进 `total_tokens`。请求结束后仍以平台用量记录与实际计费规则为准。

后端更新或重建可能使历史缓存未命中；调用方沿用 `qwen38-27b` 与现有请求格式，以实际 `usage.prompt_tokens_details.cached_tokens` 判断本次复用。

多轮对话不设固定轮数上限，限制的是每次请求的完整上下文。系统提示、工具定义、历史消息、工具结果及视觉 tokens 都占用输入预算，并须为本次输出预留空间；入口 32MiB 正文和整段历史的媒体数量限制也同时生效。调用方负责在接近预算时裁剪历史或生成摘要，保留仍需引用的图片，以及完整配对的工具调用和结果。平台不会自动丢弃历史或替客户端生成摘要；上下文超限会返回错误。缓存复用减少重复计算，不扩大上下文窗口。

八并发与 256K 是不同维度：不能据此推断八路各占满 256K 时仍全部立即执行。图片数量合规也不代表任意尺寸均能通过显存检查。语言初始化在服务就绪前完成，动态图片和视频仍可能有首次处理开销。

图片的 `image_url.url` 支持三种形式：既有 `asset://<ASSET_ID>`、公网 HTTPS 443 图片地址，以及 `...`（也支持 JPEG、WebP）。外部图片由平台自动导入临时素材，无需客户端先调用上传接口；仍受单图 16MiB、16777216 解码像素、整次请求媒体数量和上下文预算限制。URL 必须可由服务端公开访问，可带签名查询参数，不能依赖登录 Cookie 或额外鉴权头；私网、localhost、非 HTTPS 与非 443 端口拒绝。

聊天入口的整个 JSON 请求体最多 32MiB，包含 Base64 编码、所有历史和工具定义；Base64 通常比原文件大约三分之一。单图 16MiB 是解码后文件上限，多个内联图片仍可能超过请求体上限。较大素材建议使用公网 HTTPS 地址或先上传后引用 `asset://`。

视频继续按素材上传以同一 Key、同一模型上传，再在 `video_url` 中引用 `asset://<ASSET_ID>`。图文和视频工具的两跳消息格式与本页示例相同，将 `model` 改为 `qwen38-27b` 并保留原始媒体消息；使用 `stop` 时须移除媒体和工具，并设置 `stream: false`。纯文本可用字符串或标准文本片段数组，平台会规范化后转发。升级灰度期间以目标服务的公开能力为准：`toolsTextOnly: true` 时仍仅允许纯文本工具请求。

WorkBuddy 自定义模型

选择 OpenAI Chat Completions 格式，基础地址填写 `https://api.jusuanhub.com/v1`（关闭“完整 URL”）；模型 ID 为 `qwen38-27b`，填写自己的 API Key。启用工具调用与图片输入；思考可按需开启，客户端须能传递上述参数并分别处理思考与答案，不能只凭界面开关判断已生效。输出预算建议显式设置为 8192，输入预算不得超过服务公开上限。若使用完整 URL，则填写 `https://api.jusuanhub.com/v1/chat/completions`。

WorkBuddy 可直接发送 Base64 图片；无需把图片手动上传到外部图床。平台会处理已知的客户端消息附加字段、空 annotations 和工具调用的空正文，并保持工具 ID、参数、结果和历史顺序。图片应放在 user 消息中；后续对话保留原始图片消息、assistant 的 tool_calls 及对应 tool 结果。工具必须开启时也可识图，但目标服务须已开放图文工具组合。客户端取消、超时或 SSE 提前结束时，不自动重放已有用量的请求。

多个后端的总容量不是单个调用者的额度：个人用户的所有 API Key 与页面调用共享一份账号同步额度（默认 5：每秒最多发起 5 个新请求，且同时最多 5 个请求在处理，流式响应结束后才释放），不另设单 Key 门槛，服务后端容量独立生效。遇到 429 时保留错误码：ACCOUNT_REQUEST_RATE_LIMITED 表示每秒请求过多，ACCOUNT_CONCURRENCY_LIMITED 表示同时进行的请求已满，应等待已有请求完成；都要遵守 Retry-After 退避，不无限重试。需要更高额度请联系平台管理员调整。首次遇到新输入长度或视觉形状仍可能有额外处理开销，首次请求耗时不代表热态生成速度。客户端 1900 秒是等待预算示例，不是响应时间承诺；服务当前上游期限为 1860 秒。

性能比较须区分首字延迟（TTFT）、整次请求耗时和生成阶段 token/s。后台同步调用的耗时包括服务执行等待，输出吞吐按完整调用计时，不能直接和 Runtime 的纯 Decode TPS 对比。

函数工具与两跳消息

目标服务公开 tools 参数时，可发送以下请求。tool_choice 支持 auto、none、required 或指定函数对象；parallel_tool_calls: false 要求一次最多返回一个调用。平台不代替你的应用执行函数，也不会替调用方授权业务操作。

```
{
  "model": "qwen3-6-35b-a3b",
  "messages": [{"role": "user", "content": "查询杭州的天气"}],
  "tools": [{
    "type": "function",
    "function": {
      "name": "get_weather",
      "description": "查询指定城市天气",
      "parameters": {
        "type": "object",
        "properties": {"city": {"type": "string"}},
        "required": ["city"]
      }
    }
  }],
  "tool_choice": {"type": "function", "function": {"name": "get_weather"}},
  "parallel_tool_calls": false,
  "max_tokens": 256,
  "stream": false
}
```

当 finish_reason 为 tool_calls 时，解析返回的 message.tool_calls[].function.arguments JSON，校验函数名、参数和调用人的权限，再执行自己的函数。下一次请求保留原消息和完整的 assistant 消息（包括原始工具调用 ID，content 可以为 null），为每个结果追加 {"role":"tool","tool_call_id":"<RETURNED_TOOL_CALL_ID>","content":"<YOUR_TOOL_RESULT>"}。不要自己生成或改写 tool_call_id。第二跳仍是一次独立计费调用。

工具调用使用 SSE 时，按 delta.tool_calls[].index 合并 ID、函数名和参数片段，完整结束后再解析参数和执行函数。当前服务会先完整校验工具参数再分片返回，不能把它当作逐 Token 的工具参数推送。收到流内错误或提前 EOF 时，不执行不完整调用，不自动重试计费请求。

自定义停止文本

例如请求加入 "stop": ["【结束】"]。停止文本可为 Unicode；命中后正文不包含该文本，响应保留正常的结束原因与实际用量。finish_reason=length 表示达到输出预算，不能标记为完整答案。不要用 stop 截断工具参数 JSON。

/docs/models/retrieval

向量化与相关性重排

文本向量

向 POST /embeddings 提交模型别名和 input。可输入字符串或该服务允许数量的字符串数组。

```
{"model":"<EMBEDDING_MODEL_ALIAS>","input":["成都的历史","成都的天气"]}
```

成功后使用 data[].index 对齐输入，读取 data[].embedding。维度由模型决定，不在存储层写死另一模型的向量长度。批量大小和字符上限见当前参考及服务能力。

重排候选内容

先自行检索候选集合，再调用 POST /rerank。重排不会替代检索库，也不会自动抓取网页。

```
{"model":"<RERANK_MODEL_ALIAS>","query":"如何开始使用模型 API？","documents":["创建 API Key 后在服务端调用。","应用可以使用运行会话。"],"top_n":2}
```

按照响应 results[].index 回到原候选，并读取相关性分数。不要把分数当作已校准的概率；是否返回原文、候选数和 top_n 范围以模型合同为准。

失败处理

输入超长、批量过大或向量模型与重排接口混用时，应先修正输入。不自动拆成大量请求绕过限制。响应结构与错误码见对应接口参考。

OCR 与文档解析

选择能力

目标	操作	结果
从图片识别文字	POST /ocr	同步响应，不创建图片作品
将 PDF、Office 或图片解析为结构化文档	POST /documents/parse	异步任务，返回 Markdown、JSON 和可选图片 ZIP

MinerU 的公开别名为 `mineru-document-parse`。先按首次调用创建具有该服务权限的 Account API Key；在服务端使用 `https://api.jusuanhub.com/v1` 和 `Authorization: Bearer <API_KEY>`。不要把 Key 放入浏览器代码。

上传和输入限制

通过 `POST /assets/input?model=mineru-document-parse`，使用 multipart `file` 上传一个文件，读取响应的 `asset.assetId`，作为后续 `input_document_asset_id`；也可直接用 `input_document_url` 提供公网 HTTPS 地址（见直接使用公网 URL）。素材绑定原身份和服务，查询和下载使用同一 Bearer Key。

当前支持 PDF、DOCX、PPTX、XLSX、JPEG、PNG、WebP，单文件最大 256 MiB；不接受旧二进制 DOC/PPT/XLS、加密或损坏文件。以服务发现和模型卡片的当前合同为准。

输入	页数申报
PDF 全文	`pages` 为预计处理页数，范围 1–500。
PDF 部分页	使用从 1 开始的闭区间 `page_range`，如 `{"start_page":3,"end_page":8}`；按区间长度预占。
DOCX、PPTX、XLSX	必填 `pages`，申报转换后预计页数，范围 1–500；不接受 `page_range`。工作表数不等于转换页数。
JPEG、PNG、WebP	单张图片按一页处理，不接受 `page_range`。

`pages` 用于计费预占和处理上限，不能用虚假的一页绕过限额。成功后按可信 `processedPages` 结算；Office 以转换后实际解析页数为准，超过申报上限会失败。费用以活动按页报价为准。

OCR 请求

```
{"model":"<OCR_MODEL_ALIAS>","input_image_asset_id":"<ASSET_ID>","pages":1,"task":"ocr"}
```

`task` 只使用服务明确支持的枚举。OCR 输入字段不要与解析接口的 `input_document_asset_id` 混用。

文档解析请求

```
{
  "model":"mineru-document-parse",
  "input_document_asset_id":"<ASSET_ID>",
  "original_name":"document.pdf",
  "pages":12,
  "preset":"hybrid-medium",
  "language":"ch"
}
```

以 JSON 提交到 `POST /documents/parse`，携带稳定的 `Idempotency-Key`。原始文件名不含路径。预设为 `hybrid-medium` 或 `hybrid-high`，语言为 `ch` 或 `en`，中文不是 `zh`。复杂版面、公式和表格仍需人工复核。

创建后保存 `jobId`、`jobUrl` 和幂等键，按照异步任务有界轮询；遇到超时先恢复查询原任务，不自动重提。当前不提供用户取消。

下载全部结果

任务 `succeeded` 后遍历 `outputs[]`，逐个使用原样 `contentUrl` 和同一 Bearer Key 下载。输出均为文档资产；不要只保存 `outputs[0]`，也不要追加媒体水印参数。

产物	MIME	是否必需
Markdown	`text/markdown`	是
内容列表 content_list	`application/json`	是
结构化中间结果 middle_json	`application/json`	是
图片归档 images_archive	`application/zip`	有图片时可能返回

同时保存 Job 响应、各产物 MIME、字节数和 SHA-256，下载后校验。`outputs[].assetRole` 是通用资源角色，不是 `content_list` 等业务角色；业务角色见结果 `manifest.artifacts[]`，可通过摘要与输出关联。不要仅靠数组顺序区分两份 JSON。

图片 ZIP 中保留 `images/` 路径；使用 Markdown 的相对图片引用时，将图片目录放在其对应位置。结果不进入“作品与素材”，从原 Job 领取。失败任务没有可用结果；输入页数不合法、素材到期或归属不匹配时先修正原因，不猜测下载地址。

图片生成与参考图创作

一次调用的顺序

选择模型与规格 → 有参考图时先上传（或直接给公网 URL）→ 创建图片任务 → 查询终态 → 下载结果。接口为 POST /images/generations。

文生图、单图编辑、多图参考不是所有模型都支持。input_image_asset_id、input_image_asset_ids、size 和生成张数均应使用目标服务“API 接入”给出的合法字段，不要枚举清晰度和比例的所有笛卡尔积。

Qwen Image 2.1 提供文字生图、单图编辑及最多十张有序参考图，开放七种比例、两档共十四个受控尺寸；完整调用与价格说明见专属教程。

文生图请求体

下面表示请求结构。将 model 和 size 换成目标模型支持的值；不要因为其他模型接受某个尺寸就照抄。

```
{
  "model": "<MODEL_ALIAS>",
  "prompt": "明亮的产品海报，简洁背景，自然光线",
  "size": "1024x1024",
  "seed": 42
}
```

使用参考图

按 素材上传 顺序上传每张图，保存 asset.assetId。支持多图参考的模型一般使用有序数组；只支持单图的服务使用其单图字段。

```
{
  "model": "<MODEL_ALIAS>",
  "prompt": "保留人物特征，将背景改为清晨街道",
  "input_image_asset_ids": ["<FIRST_ASSET_ID>", "<SECOND_ASSET_ID>"],
  "size": "1024x1024"
}
```

此多图示例仅适用于明确支持数组输入的模型。某些模型允许零张参考图，某些编辑模式要求至少一张；数量、顺序、像素和单图大小均从对应服务合同确认。不要同时传含义冲突的单图和数组字段。

创建后如何处理

使用目标操作要求的幂等键提交。成功的 202 不是图片完成，按 异步任务 保存原任务、查询并下载。失败时不要因“没有看到图片”连续点击或重提。

同规格价格可在模型广场查看；Account API Key 不提供独立的公开估价端点，创建时按服务当前价格规则准入与结算。不要调用应用调试专用准入接口代替个人报价。

/docs/models/qwen-image-21

Qwen Image 2.1

选择输入和尺寸

公开模型别名为 `qwen-image-2.1`。使用个人 Account API Key 调用 POST /images/generations；创建返回 202 后等待任务终态。文字生图不传参考图，编辑与参考创作使用 1–10 张有序图片；提示词可按“图1”“图2”指代数组中的图片。

`prompt` 必填，最多 10,000 个 Unicode 字符。每次固定生成一张 PNG，Turbo 固定 8 步、CFG=1；不开放非空负面提示词。可省略 `steps`、`guidance_scale` 和生成张数，不要通过这些字段调整执行方式。

`seed` 可选，范围为 0–2147483647；省略时平台生成并记录实际种子。同种子用于重复同一输入，不承诺跨运行条件逐像素一致。默认输出为 1024×1024，参考图不会自动改变所选输出尺寸。

比例	高清（标准档）	超清（高分辨率档）
1:1	1024×1024（默认）	2048×2048
4:3	1184×896	2400×1792
3:4	896×1184	1792×2400
3:2	1248×832	2528×1696
2:3	832×1248	1696×2528
16:9	1376×768	2752×1536
9:16	768×1376	1536×2752

`size` 使用表中的 宽x高，例如 `1376x768`。当前只开放以上 14 个尺寸，不接受任意宽高。

文字生图

设置 JUSP_BASE_URL=https://api.jusuanhub.com/v1，JUSP_API_KEY 使用当前调用者自己的 Key。不要把 Key 写进代码仓库或浏览器页面。

```
curl --fail-with-body --max-time 120 \
  "$JUSP_BASE_URL/images/generations" \
  -H "Authorization: Bearer $JUSP_API_KEY" \
  -H 'Content-Type: application/json' \
  -H 'Idempotency-Key: qwen-image-text-001' \
  -d '{"model":"qwen-image-2.1","prompt":"清晨的山间湖泊，远山倒映，柔和自然光","size":"1376x768","seed":42}' \
  -o create.json
```

这里的幂等键代表一次不变的业务动作。重试相同动作保留原键和正文；修改提示词、图片或尺寸后使用新的业务键。

单图编辑与多图参考

每张参考图先用同一个 Key 上传，查询参数 `model` 指向本模型：

```
curl --fail-with-body --max-time 120 \
  "$JUSP_BASE_URL/assets/input?model=qwen-image-2.1" \
  -H "Authorization: Bearer $JUSP_API_KEY" \
  -F 'image=@reference.png' -o upload.json
```

保存响应的 `asset.assetId`。单图编辑请求如下；将占位符换成已上传且当前身份可用的 Asset ID：

```
{
  "model": "qwen-image-2.1",
  "prompt": "保留图1的主体，将背景改为浅蓝色摄影棚。",
  "size": "2048x2048",
  "seed": 42,
  "input_image_asset_ids": ["<ASSET_1>"]
}
```

多图使用同一接口和字段，按提示词中的引用顺序逐项传入。两张示例：

```
{
  "model": "qwen-image-2.1",
  "prompt": "将图1的人物放入图2的场景，保留人物特征，统一光线。",
  "size": "1376x768",
  "input_image_asset_ids": ["<ASSET_1>", "<ASSET_2>"]
}
```

十张参考图仍是一个任务，数组完整保留，不拼图或静默截断：

```
{
  "model": "qwen-image-2.1",
  "prompt": "按图1到图10顺序，以各图主体为元素创作一张协调的展览海报。",
  "size": "2048x2048",
  "input_image_asset_ids": [
    "<ASSET_1>", "<ASSET_2>", "<ASSET_3>", "<ASSET_4>", "<ASSET_5>",
    "<ASSET_6>", "<ASSET_7>", "<ASSET_8>", "<ASSET_9>", "<ASSET_10>"
  ]
}
```

图片限 PNG、JPEG、静态 WebP；单张最大 16 MiB、最长边 5760、最多 40,000,000 像素、宽高比不超过 200:1。平台还校验参考图与输出图的联合预算，极端比例或复杂组合即使不足十张也可能被拒绝。传递原文件的 MIME，不通过改扩展名绕过校验；不接受任意 URL、文件路径或直接 base64。Asset 有效期与归属规则见素材输入。

查询、下载与失败恢复

保存创建响应中的 `jobId` 和 `jobUrl`，携带原 Key 查询；等价查询路径为 `GET /jobs/{jobId}?model=qwen-image-2.1`。`queued`、`running` 继续等待，`succeeded` 后读取 `outputs[0].contentUrl` 并携带原 Key 下载。输出为 PNG；下载失败只重试领取原结果，不重新生成。

完整的有界轮询、同源 URL 检查和下载示例见 API Key 异步媒体完整调用与任务恢复。`failed` 是业务终态，读取错误码后再由用户决定下一次调用。`429` 按 `Retry-After` 退避；超时或断网不表示服务端未受理，先恢复原任务。当前不支持主动取消运行中的生成。

提示词优化与费用

提示词优化为**手动可选**。使用独立优化入口得到 `optimization.optimizedPrompt` 文本，用户确认后再将其放入本次生成请求的 `prompt`；保留已选 `size`，不会默认自动生成。优化遵循 Qwen 官方改写规则，模型正文直接返回文本，不要求 H3 的结构化 JSON 或修复流程。

带图优化按原顺序转交全部参考图；优化模型有自己的输入上限，超限错误原样反馈，不会丢图。生成支持十图不代表当前优化模型也能接受十图；优化失败可由用户保留原文继续创建生成任务。优化单独按所用 Qwen3.8 模型计费，和后续生图是两次独立调用。

高清标价 **8 积分/张**，超清标价 **16 积分/张**。适用全局五折时为 **4/8 积分/张**；具体折扣和最终扣费以调用者当时有效价格及回执为准。参考图数量不增加输出张数；输入不合规不应反复提交，受理、预占、释放与实际结算是不同状态，不能把失败或账务待确认当作“已扣零积分”。

MiniMax H3 四种视频模式

需要外部素材 URL 或 New API 中转接入时，使用 H3 V2 兼容入口。下文为原生 V1 接口。

先选择模式

H3 通过 POST /media/generations 创建视频任务。公开模型别名以 Portal 详情为准；不要使用服务实例 ID。

模式	generationMode	输入
文生视频	`t2v`	提示词，不传素材
图生视频	`i2v`	`input_image_asset_id`
首尾帧	`first_last_frame_video`	`input_image_asset_id` 和 `end_image_asset_id`
全能参考	`universal_reference_video`	有序 `referenceInputs`，至少一张图片或一个视频

当前 H3 合同支持 `768p` 和 `544p` 两档、5–15 秒整数时长、24 FPS 音画同步 MP4。每档只接受服务详情列出的六种固定画布；可用 `aspectRatio` 选择比例，也可直接使用 `outputSizeCode`。不要传任意像素、内部帧数、采样步数或插帧参数。

档位	支持画布（按比例）
`768p`	`21:9` 1536×672、`16:9` 1344×768、`4:3` 1024×768、`1:1` 768×768、`3:4` 768×1024、`9:16` 768×1344
`544p`	`21:9` 1280×544、`16:9` 960×544、`4:3` 736×544、`1:1` 544×544、`3:4` 544×736、`9:16` 544×960

不传新增字段的历史 `768p` 请求继续按原合同处理。全能参考前 6 张图片按基础单价计费，第 7 张起每增加一张图片，单位积分单价增加 10 积分。

创建请求

先按 快速开始 设置三个环境变量。将下列 JSON 保存为 `request.json`，替换 `model`。相同业务请求使用稳定的幂等键。

```
{
  "model":"<MODEL_ALIAS>",
  "generationMode":"t2v",
  "prompt":"雨后街道，镜头缓慢前移，人物撑伞走过，保留自然雨声",
  "resolutionTier":"768p",
  "orientation":"landscape",
  "aspectRatio":"16:9",
  "seconds":5
}
```

```
curl --fail-with-body --max-time 60 "$JUSP_BASE_URL/media/generations" \
  -H "Authorization: Bearer $JUSP_API_KEY" \
  -H 'Content-Type: application/json' \
  -H 'Idempotency-Key: your-stable-business-request-id' \
  --data-binary @request.json --output create.json
```

图生视频完整请求

先上传图片，保存首图 ID（也可把 `input_image_asset_id` 换成公网地址 `input_image_url`，见直接使用公网 URL），再将以下完整 JSON 保存为 `request.json`，沿用上面的提交命令：

```
{"model":"<MODEL_ALIAS>","generationMode":"i2v","prompt":"保持首图人物和场景，镜头缓慢推进","input_image_asset_id":"
<IMAGE_ASSET_ID>","resolutionTier":"768p","orientation":"landscape","seconds":5}
```

首尾帧完整请求

两张图均需提前上传，不把本地路径放进 JSON：

```
{"model":"<MODEL_ALIAS>","generationMode":"first_last_frame_video","prompt":"从首图自然过渡到尾图，人物动作连
贯","input_image_asset_id":"<START_IMAGE_ASSET_ID>","end_image_asset_id":"
<END_IMAGE_ASSET_ID>","resolutionTier":"768p","orientation":"landscape","seconds":5}
```

全能参考

以下是全能参考完整请求；三类素材均需提前上传。只用图片时删除视频和音频两项，不填空 ID：

```
{
  "model":"<MODEL_ALIAS>",
  "generationMode":"universal_reference_video",
  "prompt":"保持参考图的主体，参考视频的运镜与参考音频的节奏",
  "resolutionTier":"768p",
  "orientation":"landscape",
  "seconds":5,
  "referenceInputs":[
    {"role":"reference_image","mediaType":"image","assetId":"<IMAGE_ASSET_ID>"},
    {"role":"reference_video","mediaType":"video","assetId":"<VIDEO_ASSET_ID>"},
    {"role":"reference_audio","mediaType":"audio","assetId":"<AUDIO_ASSET_ID>"}
  ]
}
```

将完整请求保存为 `request.json`，沿用创建命令。参考顺序有意义：图片、视频、音频分别有序，不重复同一个素材。提示词可明确每项素材的用途。

限制	当前合同
图片	0–9 张；存在视频时最多 8 张
视频	0–1 个
音频	0–3 段，每段 2–15 秒，总时长不超过 15 秒
总数	不超过 12；至少一个视觉素材
提示词	不超过 7000 个字符

V1 与 V2 使用同一角色素材限制，该表统一列出格式、字节、解码像素、尺寸、时长和宽高比边界。首尾帧与全能参考图片的文件大小、长边上限不同，首尾帧也不套用全能参考的宽高比限制；请按所选模式和角色校验。

文件字节数与媒体时长须分别满足限制；请求 `seconds` 的5–15秒是输出时长。MOV 音频参考必须按 `audio` 上传身份处理，且有有效音轨，不占视频参考名额；参考视频自带音轨不计入独立音频总长。上传成功只表示已暂存，通过上传与错误处理说明了解最终按角色校验的要求。

默认智能优化与费用

只需创建一次生成任务，无需额外调用优化接口。使用 Account API Key 创建 H3 四种模式任务时，`optimizePrompt` 缺省为 `true`。平台先优化提示词，成功后才生成视频；创建仍返回异步 Job。只接受 JSON 布尔值，`null`、字符串和数字会被拒绝。其他模型不会因省略字段自动开启优化。

平台尚未启用自动优化能力时，H3 缺省或 `optimizePrompt: true` 请求返回 `503 prompt_optimization_capability_unavailable`。此时不会按原文自动生成；需要直接使用已确认文本时显式传 `false`，仍需满足正常生成审核与规格要求。默认优化示例以平台已启用该能力为前提。

如果已准备好最终提示词，或已独立调用优化接口并人工采用结果，请显式关闭，避免重复优化及费用：

```
{"model":"<MODEL_ALIAS>","generationMode":"t2v","prompt":"人工确认的最终提示
词","optimizePrompt":false,"resolutionTier":"768p","orientation":"landscape","seconds":5}
```

false 的优化调用和优化费用均为零，视频生成仍按正常价格计费并经过内容安全审核。开启时，优化和视频生成分别计费；最多两次常规优化尝试，每次完整超长候选最多再压缩一次，合计最多四次优化模型调用，按实际用量计费。已经执行的优化调用即使最终失败也可能收费。

已知可重试技术失败会重试一次，再失败时 Job 返回 failed 和 errorCode: prompt_optimization_failed，保留原文且不自动生成。安全拒绝、需要澄清和候选拒绝不通过重试寻找放行。完整费用、人工采用流程及未知结果恢复见优化教程。个人创作保持人工采用流程；App 的 H3 默认优化及关闭方式见内部应用开发中的平台服务教程。

自动优化的错误会区分原因：准备超时为 prompt_optimization_timeout；内容审核未通过为 content_safety_blocked；需要补充说明为 content_safety_requires_clarification；优化稿被拒绝为 prompt_optimization_candidate_rejected；审核服务暂不可用为 content_safety_unavailable。这些错误发生于 failureStage: prompt_optimization，安全相关原因的 phase 为 content_safety。技术失败继续使用 prompt_optimization_failed。响应说明视频生成尚未开始，优化费用仍按实际调用结算。

查询 HTTP 200 仅表示查询成功，必须检查正文 status、errorCode、errorMessage；有 details 时可显示受控的审核原因和操作指导。retryable: false 表示不要自动新建付费任务。恢复查询保留原失败原因。

查询与恢复

202 返回后按 任务查询 继续。用原 Key 访问原 jobUrl，完成后从 outputs[] 下载。客户端等待结束不取消任务，不要换键重新生成。

/docs/models/video/h3-v2

H3 V2 直接调用

从自己的服务端直接调用平台，请按本文操作。中转站管理员请阅读 New API 接入教程；逐字段定义见 V2 接口参考。

完成第一次调用

1. 在平台创建 API Key，在自己的服务端设置 JUSP_API_KEY。
2. 从下方四种模式中选一份完整 JSON，保存为 request.json；素材 URL 换成实际公网地址。
3. 使用下面的创建命令，保存 create.json。同一业务重试保持正文和幂等键；新业务换键。
4. 下载 Python 轮询下载脚本，与 create.json 放在同一目录，运行 python3 h3-v2-poll.py。脚本只查询原任务，不创建新任务。

```
curl --fail-with-body --max-time 120 \
  'https://api.jusuanhub.com/v2/video_generation' \
  -H "Authorization: Bearer $JUSP_API_KEY" \
  -H 'Content-Type: application/json' \
  -H 'Idempotency-Key: your-stable-business-request-id' \
  --data-binary @request.json --output create.json
```

创建返回 task_id 只代表受理；最终成功、失败与下载方式见 查询、下载与失败处理。本页不会执行请求或保存密钥。

接口与支持范围

本接口是 MiniMax H3 的兼容子集，使用平台 Account API Key。原 V1 媒体接口 保持不变；V2 无需先上传素材获取 assetId。

操作	地址
创建	POST https://api.jusuanhub.com/v2/video_generation
查询	GET https://api.jusuanhub.com/v2/query/video_generation/{task_id}
下载	查询成功后使用 task.content.url，支持 GET、HEAD、Range

创建和查询使用 Authorization: Bearer <API_KEY>。模型名为 MiniMax-H3，清晰度仅 768P，时长 duration 为 5–15 秒整数。2K、4 秒、回调、多参考视频和 aigc_watermark: true 不支持。

所有模式统一固定画幅

每个新请求都必须显式填写 ratio，只接受下表六个值。 不传、adaptive、其他比例或自定义宽高都会被拒绝。素材原始尺寸不会改变所选画幅。

ratio	输出宽×高
21:9	1536×672
16:9	1344×768
4:3	1024×768
1:1	768×768
3:4	768×1024
9:16	768×1344

比例名称表示平台预设档位，精确像素以表格及查询返回的 width、height 为准。

> **画幅小提示：** 生成视频使用所选固定画幅。建议上传与目标画幅比例一致的图片或参考视频；比例不一致可能造成画面拉伸、压缩、裁剪或构图变化。可在提交前自行裁剪到相同比例；首尾帧建议使用一致的画幅。全能参考用于引导生成，不保证逐像素保留原构图。

用素材角色明确选择模式

content 包含一项 text，以及所需素材项。图片数量不能代替模式选择：一张 first_frame 是图生，一张 reference_image 是全能参考。

模式	素材角色
文生	只有 text
图生	一张 image_url，role 为 first_frame
首尾帧	两张 image_url，role 分别为 first_frame、last_frame
全能参考	reference_image、reference_video、reference_audio，至少一项视觉素材

不要混合首尾帧角色和全能参考角色。省略图片 role 仅为兼容而默认首帧，两张图片都省略不会自动成为首尾帧。

四种模式的完整请求

以下每份 JSON 均可独立保存为 request.json，使用开头的创建命令。替换示例素材 URL 为你可公开读取的实际地址；示例域名没有可用素材。content 恰好包含一项非空 text，提示词最多 7000 字符。不传不支持的字段，名称和大小写须精确。

文生视频

```
{
  "model": "MiniMax-H3",
  "resolution": "768P",
  "duration": 5,
  "ratio": "16:9",
  "content": [
    {
      "type": "text",
      "text": "保持场景主体，镜头缓慢推进，动作自然，保留环境声"
    }
  ]
}
```

图生视频

```
{
  "model": "MiniMax-H3",
  "resolution": "768P",
  "duration": 5,
  "ratio": "16:9",
  "content": [
    {
      "type": "text",
      "text": "保持场景主体，镜头缓慢推进，动作自然，保留环境声"
    },
    {
      "type": "image_url",
      "role": "first_frame",
      "image_url": {
        "url": "https://your-public-host.example/first.png"
      }
    }
  ]
}
```

首尾帧

```
{
  "model": "MiniMax-H3",
  "resolution": "768P",
  "duration": 5,
  "ratio": "16:9",
  "content": [
    {
      "type": "text",
      "text": "保持场景主体，镜头缓慢推进，动作自然，保留环境声"
    },
    {
      "type": "image_url",
      "role": "first_frame",
      "image_url": {
        "url": "https://your-public-host.example/first.png"
      }
    },
    {
      "type": "image_url",
      "role": "last_frame",
      "image_url": {
        "url": "https://your-public-host.example/last.png"
      }
    }
  ]
}
```

全能参考

```
{
  "model": "MiniMax-H3",
  "resolution": "768P",
  "duration": 5,
  "ratio": "16:9",
  "content": [
    {
      "type": "text",
      "text": "保持场景主体，镜头缓慢推进，动作自然，保留环境声"
    },
    {
      "type": "image_url",
      "role": "reference_image",
      "image_url": {
        "url": "https://your-public-host.example/reference.png"
      }
    },
    {
      "type": "video_url",
      "role": "reference_video",
      "video_url": {
        "url": "https://your-public-host.example/reference.mp4"
      }
    },
    {
      "type": "audio_url",
      "role": "reference_audio",
      "audio_url": {
        "url": "https://your-public-host.example/reference.wav"
      }
    }
  ]
}
```

外部素材与费用

- 素材地址必须是公网 HTTPS、端口 443；不支持内网、回环地址、URL 用户名密码或附加鉴权头。图片另支持 PNG/JPEG/WebP base64 data URL；视频使用 MP4，音频和各角色限制见下表。跳转后的地址同样必须满足公网限制；不接受 URL fragment。
- 无视频最多 9 张图，有视频最多 8 张图；最多 1 个视频、3 段音频，音频总长不超过 15 秒；总引用不超过 12 项，纯音频参考不支持。
- JSON 最多 64 MiB，素材下载总量最多 128 MiB；单次下载最多 30 秒，总下载上下文最多 90 秒。签名 URL 必须在提交摄入期间有效。
- 提示词优化默认开启，与视频分别计费。直接 V2 请求使用 JSON `"optimizePrompt": false` 关闭，`prompt_optimizer` 为别名，两者同时提供必须一致。字符串 `"false"` 无效。
- 素材是临时任务输入，不加入个人素材库。中心输入一小时到期，在途任务保护其字节；执行缓存保留现有终态后 24 小时恢复窗口，随后按清理周期回收。不能承诺所有副本一小时内物理删除。
- 长期记录保留脱敏来源和媒体摘要，不保存完整签名 URL。临时素材清理后不可直接复用；生成视频遵循原结果生命周期，请及时下载。结果 URL 最长成功后七天有效，资产或权限提前失效时同步失效。

各角色素材限制

角色	格式	单文件上限	尺寸或时长
first_frame / last_frame	PNG、JPEG、WebP	16 MiB	长边≤8192，解码像素≤4000万
reference_image	PNG、JPEG、WebP	30 MiB	长边≤5760，解码像素≤4000万，宽高比0.4–2.5
reference_video	MP4	50 MiB	每段2–15秒，长边≤5760，宽高比0.4–2.5；须有可解码视频轨
reference_audio	WAV、MP3、FLAC、M4A、AAC、OGG，或以音频身份解析的 MOV	15 MiB	每段2–15秒，合计≤15秒

上述时长和宽高比范围均包含端点；宽高比为素材宽度除以高度，解码像素为宽度乘以高度。首尾帧不套用全能参考的0.4–2.5宽高比限制。参考视频自带音轨不占独立参考音频名额；MOV 按音频身份解析时须有有效音轨，不占参考视频名额。

文件大小与时长是独立限制，必须同时满足。`1 MiB = 1,048,576` 字节；相同时长的文件会因编码、码率、分辨率不同而大小不同，不能由字节数换算是否满足2–15秒。请求 `duration` 的5–15秒是生成视频时长，不是参考素材的时长范围。

平台检查文件真实内容，不以扩展名代替格式检测。请求总量和素材数量还须同时满足上文限制；不能把单文件上限简单乘以素材数量。V1 与 V2 使用同一角色素材限制。最新角色规格可从 `GET /v1/services` 对应 H3 的 `mediaSpecContract.inputAssets.roleLimits` 查询，按 `generationMode` 和 `role` 匹配；这只是发现规格，V2 提交仍使用 URL。

默认优化失败时任务会失败，不自动退回原文生成。已经执行的优化即使失败也可能计费；关闭优化不跳过内容安全审核。优化重试、费用和失败分类见提示词优化教程，其 V1 响应字段不能直接套到 V2。

查询、下载与失败处理

创建后先持久保存 `task_id`、原请求和幂等键。查询使用原平台 API Key:

```
TASK_ID=$(jq -er '.task_id' create.json)
curl --fail-with-body --max-time 60 \
  -H "Authorization: Bearer $JUSP_API_KEY" \
  "https://api.jusuanhub.com/v2/query/video_generation/$TASK_ID" --output task.json
```

task.status	处理
queued / running	继续查询原任务；建议间隔5秒并设置总等待期限
succeeded	读取 task.content.url 下载；检查实际 width / height
failed	停止，读取 task.error.code / message；不自动新建任务

task.status	处理
cancelled	停止；本兼容入口未开放取消接口

成功查询响应示意（用量为示意值）：

```
{"task":{"id":"job_example","model":"MiniMax-
H3","status":"succeeded","created_at":1800000000,"updated_at":1800000060,"resolution":"768P","duration":5,"ratio":"16:9","widt
h":1344,"height":768,"task_type":"generation","modality":"video","content":
{"url":"https://api.jusuanhub.com/v2/video_generation/results/asset_example/content?ticket=REDACTED"},"usage":
{"input_seconds":0,"input_image_count":0,"output_seconds":5.167,"total_seconds":5.167}}}
```

任务失败时可能返回 HTTP 200 和 `{"task":{"status":"failed","error":{"code":"prompt_optimization_failed","message":"..."}}}`；这里省略了其他任务字段。HTTP 请求本身失败则是 `{"type":"error","error":{"type":"bad_request_error","code":"unsupported_request","message":"...","http_code":"400"},"request_id":"..."}`。

仅成功后下载，URL 已带短期票据，**不需要额外 Bearer，也不要把 API Key 转发到下载地址**：

```
# 确认 task.status 为 succeeded 后执行
RESULT_URL=$(jq -er '.task.content.url' task.json)
curl --fail-with-body --max-time 300 "$RESULT_URL" --output result.mp4
# 元信息和断点读取同样使用完整票据 URL
curl --fail-with-body --head "$RESULT_URL"
curl --fail-with-body --range 0-127 "$RESULT_URL" --output first-128-bytes.bin
```

下载 URL 是敏感凭证，不写入公开日志。票据绑定原 API Key 的状态、权限和结果资产；Key 撤销或资产失效会让 URL 提前失效。过期后不能靠重新查询无限延长有效期。

可下载完整 Python 轮询下载示例，放在 `create.json` 同目录，安全提供 `JUSP_API_KEY` 后运行 `python3 h3-v2-poll.py`。脚本最多等待30分钟，对少量临时读取错误退避，下载预算1 GiB；从不提交新任务。等待结束后重跑脚本继续原任务即可。V1 的 jobUrl、outputs[] 和鉴权下载教程不适用于此响应。

情况	恢复方式
400 / 413 / 415	修正参数、URL、素材或大小限制；未受理时没有 task_id
401 / 403 / 404	核对 Key、权限、模型或任务归属；不自动切换身份
402	核对账户余额和费用；不要循环重试
409	同键参数冲突；原业务恢复原正文，新业务使用新键
查询429 / 502 / 503 / 504或网络失败	有界退避后继续查询原 task_id；不因此重新生成
创建超时、断网、服务错误	结果可能未知；保留原 Key、正文和幂等键重放，不能换键或盲目换渠道
查询410 result_window_closed，或下载票据失效	结果窗口或权限已失效；不自动重新付费生成

成功 `usage.input_seconds` 为参考视频时长，`input_image_count` 为输入图片数；`output_seconds` 为探测的实际产物时长，可能与请求 `duration` 略有差异；`total_seconds` 为参考视频与产物时长之和，不包含参考音频。它们不是金额，优化用量也不包含在其中；平台费用以实际账单为准，中转站另行配置零售计费。

New API 配置与请求示例

中转站接入已独立整理到 New API 接入教程，包含渠道配置、四种完整请求和专用轮询下载脚本。平台 V2 与 New API 的令牌、查询状态和下载鉴权不同，请勿混用。

接口字段见 V2 创建参考、V2 查询参考及 完整 OpenAPI。

已接受历史任务

已接受的旧 adaptive 任务继续使用原冻结画布，查询、下载和同键同请求重放保持可用；不会重新下载素材、改尺寸或重新生成。新请求适用六档规则。同一固定任务更换 ratio 必须使用新的业务幂等键；旧键修改参数会冲突。

`/docs/models/video/h3-newapi`

H3 · New API 中转站接入

本文面向 New API 中转站管理员。New API 是由你部署的接口转发和计费服务；平台 H3 是其上游。无需 New API 的开发者请使用 平台 V2 直接调用。

分清两段调用

调用方向	地址	使用的密钥
New API → 平台 H3	`https://api.jusuanhub.com`	平台 Account API Key，仅填渠道上游配置
客户程序 → 你的 New API	你的站点地址	该 New API 站点发给客户的令牌

不要把平台 Key 发给中转站客户。 下文所有客户请求均发往自己的 New API 站点，不能把 `/v1/tasks/hailuo` 拼到平台地址。

配置上游渠道

已验证官方 New API 源码版本 `1d4328e97417a043a161a0dd30a5b129be3ace49` 的 Hailuo 插件 1.2.0。使用 **Task Plugin（任务插件，渠道类型61）** 并选择 `hailuo`；不要将旧版内置海螺渠道与任务插件混用。基础地址填 `https://api.jusuanhub.com`，不要追加 `/v1`；Key 填平台 Account API Key，模型映射使用 `MiniMax-H3`。客户其他版本或分支需另行核对。

接入顺序:

1. 在支持任务插件的 New API 版本中加载并启用官方 `hailuo` 插件，新增任务插件渠道；插件设置对应 `task_plugin_key: "hailuo"`。
2. 填写上述基础地址和平台 Key，添加模型 `MiniMax-H3`，并将渠道分组与客户令牌可用分组对齐。
3. 在该版本的模型计费配置中为 `MiniMax-H3` / `hailuo` 配置零售价格或计费表达式，并确保客户账户有额度。不要将测试费率或平台账单金额直接当作你的零售价格。
4. 用 New API 客户令牌向下述任务接口提交，再通过 New API 查询和下载。UI 名称随版本变化，缺少任务插件入口的版本不能照此接入。

必须显式提供 `metadata.ratio`。 该插件遇到视觉素材且未指定比例时会自动发送 `adaptive`，平台会拒绝这个新请求。插件已有固定比例透传能力。

插件能力边界

能力	官方 Hailuo 1.2.0
四种视频模式、外部 URL 素材	支持，按下文明确传递素材角色
六档固定画幅	支持；必须填写 `metadata.ratio`
提示词优化	默认开启、单独计费；不透传关闭开关
稳定的上游幂等键	默认不透传；New API 任务 ID 不能替代平台幂等键

需要关闭优化时，须由接入方适配插件透传，不能仅在客户 JSON 添加 `optimizePrompt: false` 就认为已生效。创建超时或结果未知时，先核对原任务；不要自动重新提交或切换渠道，避免重复生成与计费。

四种模式的客户请求

每份 JSON 可独立保存为 `newapi-request.json`。模型固定 `MiniMax-H3`，清晰度 `768P`，时长5–15秒整数；比例为 `21:9`、`16:9`、`4:3`、`1:1`、`3:4`、`9:16`。

素材必须使用有效的公网 HTTPS URL。示例域名没有实际素材；格式、数量、大小和临时存储规则统一见 平台素材限制，单文件时长、尺寸和宽高比见各角色素材限制，精确输出尺寸见 画幅表。参考视频的2–15秒与输出的5–15秒须分别校验；文件字节数不能换算为秒数。

文生视频

```
{
  "model": "MiniMax-H3",
  "prompt": "保持主体，镜头缓慢推进，动作自然，保留环境声",
  "duration": 5,
  "metadata": {
    "resolution": "768P",
    "ratio": "16:9"
  }
}
```

图生视频

```
{
  "model": "MiniMax-H3",
  "prompt": "保持主体，镜头缓慢推进，动作自然，保留环境声",
  "duration": 5,
  "metadata": {
    "resolution": "768P",
    "ratio": "16:9",
    "first_frame_image": "https://your-public-host.example/first.png"
  }
}
```

首尾帧

```
{
  "model": "MiniMax-H3",
  "prompt": "保持主体，镜头缓慢推进，动作自然，保留环境声",
  "duration": 5,
  "metadata": {
    "resolution": "768P",
    "ratio": "16:9",
    "first_frame_image": "https://your-public-host.example/first.png",
    "last_frame_image": "https://your-public-host.example/last.png"
  }
}
```

全能参考

```
{
  "model": "MiniMax-H3",
  "prompt": "保持主体，镜头缓慢推进，动作自然，保留环境声",
  "duration": 5,
  "metadata": {
    "resolution": "768P",
    "ratio": "16:9",
    "content": [
      {
        "type": "text",
        "text": "保持主体，镜头缓慢推进，动作自然，保留环境声"
      },
      {
        "type": "image_url",
        "role": "reference_image",
        "image_url": {
          "url": "https://your-public-host.example/reference.png"
        }
      },
      {
        "type": "video_url",
        "role": "reference_video",
        "video_url": {
          "url": "https://your-public-host.example/reference.mp4"
        }
      },
      {
        "type": "audio_url",
        "role": "reference_audio",
        "audio_url": {
          "url": "https://your-public-host.example/reference.wav"
        }
      }
    ]
  }
}
```

全能参考中删除不使用的素材项，至少保留一项视觉素材；不要将首尾帧字段与 `metadata.content` 的参考角色混用。图片数量不决定模式，素材角色才决定模式。

创建、轮询与下载

在服务端安全设置 `NEW_API_ORIGIN`（你的站点基础地址，不带 `/v1`）和 `NEW_API_TOKEN`（客户令牌），不要把令牌写入文件或公开日志。

```
curl --fail-with-body --max-time 120 "$NEW_API_ORIGIN/v1/tasks/hailuo" \
  -H "Authorization: Bearer $NEW_API_TOKEN" \
  -H 'Content-Type: application/json' \
  --data-binary @newapi-request.json --output newapi-create.json
```

保存返回的 New API `task_id`。下载 New API 专用 Python 轮询下载脚本，与 `newapi-create.json` 放在同一目录，运行 `python3 h3-newapi-poll.py`。需要 Python 3；无需额外依赖。脚本最多等待30分钟，限制连续读取重试，下载预算 1 GiB；只读取原任务，绝不重新创建。

操作	New API 客户接口	鉴权
查询	`GET /v1/tasks/{task_id}`	客户令牌 Bearer
列出产物	`GET /v1/tasks/{task_id}/artifacts`	客户令牌 Bearer
下载视频	`GET /v1/tasks/{task_id}/artifacts/video/content`	客户令牌 Bearer

查询状态与恢复

顶层 `status`	客户端处理
`NOT_START`、`SUBMITTED`、`QUEUED`、`IN_PROGRESS`	继续查询同一任务，建议间隔5秒并限制总等待时间
`SUCCESS`	下载视频；检查文件实际画幅和时长
`FAILURE`	停止轮询，检查 `fail_reason`；不自动重新付费生成
未知状态	停止并核对版本，不能当作成功

New API 查询示意：

```
{"task_id":"task_example","status":"SUCCESS","progress":"100%","fail_reason":""}
```

平台 V2 的 `task.status`、小写状态和票据下载规则不适用于上述客户接口。客户下载必须携带 New API 令牌。不要把上游任务 ID 替换成 New API 任务 ID，反之亦然。

401/403先检查客户令牌与分组，404检查任务归属和版本，余额不足时检查客户额度与零售计费。查询429/502/503/504可有界退避；创建失败或超时先区分未受理与结果未知。脚本等待结束不代表任务已取消，重新运行脚本会继续查询原任务。

接入验收清单

1. 渠道能识别 `MiniMax-H3`，模型、分组、客户额度与零售计费配置正确。
2. 用一份请求完成创建 → 查询 `SUCCESS` → 携带客户令牌下载；保存原任务 ID。
3. 核对实际视频尺寸符合所选画幅，并检查声音和画面。需要其他模式时分别替换素材验证。
4. 确认默认优化与视频分别计费；未知提交结果不会触发自动换渠道重试。

平台逐字段定义见 V2 接口参考。该参考描述上游平台接口，不是 New API 客户协议。

Wan 角色动画与角色替换

Wan2.2-Animate-14B 通过 POST /media/generations 接入。它不是普通提示词驱动文生视频：两种模式都使用角色图和源视频。

选择 characterMode

模式	值	用途
动作迁移	`motion_transfer`	把视频动作迁移到人物图
角色替换	`character_replace`	按服务能力替换视频角色

先分别通过 `image` 和 `video` 字段上传，保存两个素材 ID。

```
{
  "model":"<MODEL_ALIAS>",
  "characterMode":"motion_transfer",
  "input_image_asset_id":"<IMAGE_ASSET_ID>",
  "input_video_asset_id":"<VIDEO_ASSET_ID>",
  "resolutionTier":"720p",
  "speedMultiplier":1.5
}
```

角色替换使用以下独立请求，其余输入准备相同：

```
{"model":"<MODEL_ALIAS>","characterMode":"character_replace","input_image_asset_id":"
<IMAGE_ASSET_ID>","input_video_asset_id":"<VIDEO_ASSET_ID>","resolutionTier":"720p","speedMultiplier":1}
```

该模型不传提示词、任意秒数、画幅、FPS 或普通视频服务的 `enableRife`。按 快速开始 配置环境，将所选完整 JSON 保存为 `request.json`：

```
curl --fail-with-body --max-time 60 "$JUSP_BASE_URL/media/generations" \
  -H "Authorization: Bearer $JUSP_API_KEY" \
  -H 'Content-Type: application/json' \
  -H 'Idempotency-Key: character-shot-001' \
  --data-binary @request.json --output create.json
```

处理速度与价格

当前开放 480p / 720p；720p 可选 1x 或 1.5x。1.5x 的基础报价按对应速度系数调整，不是免费加速。服务是否接受该组合、输入时长和输出帧上限均以当前合同为准。

清晰度	处理速度	当前有效帧上限	分段上限
480p	1x	305	4
720p	1x	305	4
720p	1.5x	609	8

每段按 77 帧的合同处理；30 FPS 下上表约对应 10.2 秒和 20.3 秒，但不得用这个近似值替代服务对有效帧的校验。480p 不支持 1.5x。两种角色模式均需满足当前服务输入限制。

上传与解码、排队等耗时不与速度倍数简单成反比；不承诺整条请求必然快 1.5 倍。最终按实际请求规格和生效价格规则计费。

完成链路

使用服务要求的幂等键创建后，按照 异步任务 查询和下载；不要复用 H3 的全能参考数组作为本服务输入。

/docs/models/video/other

选择其他视频能力

这些能力共享 POST /media/generations，但不共享完全相同的输入。只选择当前模型列表中已开放、且你的 Key 可以调用的服务。

按目标选择教程

你的目标	所需输入	完整教程
从提示词或首图生成视频	提示词，可选首图	普通视频生成
使用多图或已有视频作为参考	随模式提供图片、视频或组合	Bernini 四种模式
提升已有视频清晰度	一个 MP4 视频，不需要提示词	视频超分
让人物嘴型匹配音频	人物视频和驱动音频，不需要提示词	X-Dub 口型同步

H3 和 Wan 角色动画分别见 H3 四种模式 与 Wan 动作迁移和角色替换。教程描述接入合同，不代表每个账号都已获得全部能力；先完成 服务发现。不要把一个模型的参数集合直接套给另一个模型。

统一恢复方式

所有异步分支保存原任务并 查询终态、下载结果。失败后的恢复以当前错误信息为准，不跨模型复用素材，不因超时就重新生成。

/docs/models/video/standard

普通文生与图生视频

本页针对普通视频生成服务，不适用于 H3、Wan 角色动画、口型同步或视频超分。先按 快速开始 设置 `JUSP_BASE_URL`、`JUSP_API_KEY`、`JUSP_MODEL`，再读取 服务能力。

选择模式与输入

`t2v` 使用提示词；`i2v` 还需要一张首图。图生模式先通过 输入素材上传 的 `image` 字段上传图片，保存 `asset.assetId`。清晰度、画幅、时长和模式必须是当前服务开放的组合；下面的 480p、横屏、5 秒仅适用于支持该组合的服务。

```
{"model":"<MODEL_ALIAS>","generationMode":"t2v","prompt":"海边日落，固定镜头，海浪缓慢向岸边推
进","resolutionTier":"480p","orientation":"landscape","seconds":5,"seed":42}
```

图生视频请求:

```
{"model":"<MODEL_ALIAS>","generationMode":"i2v","prompt":"保持首图主体，镜头缓慢推进","input_image_asset_id":"
<IMAGE_ASSET_ID>","resolutionTier":"480p","orientation":"landscape","seconds":5,"seed":42}
```

`enableRife` 只在普通 Wan 服务明确支持插帧时传入。不应因其他视频模型接受某字段，就向当前模型传递同一字段。不要自行传入内部步数、任意 FPS 或执行批大小。

提交并完成任务

将一个完整 JSON 保存为 `request.json`，替换模型别名和素材 ID:

```
curl --fail-with-body --max-time 60 "$JUSP_BASE_URL/media/generations" \
  -H "Authorization: Bearer $JUSP_API_KEY" \
  -H 'Content-Type: application/json' \
  -H 'Idempotency-Key: video-shot-001' \
  --data-binary @request.json --output create.json
```

HTTP 202 仅表示受理。按 等待任务并下载结果 查询原 `jobUrl`，成功后下载 `outputs[]`；按响应 `mimeType` 保存文件。失败或本地等待超时不会自动创建替代任务。

参数失败时怎样恢复

400 表示模式、素材或规格不合法，先修正请求；401/403 检查 Key 和服务范围；429 按 `Retry-After` 等待。修改正文是新的业务请求，不得在原幂等键下换参数。价格依据当前服务的基础单价、计费单位及规格，不用任务执行耗时估算生成费用。

/docs/models/video/bernini

Bernini 参考视频生成

首先完成 快速开始 与 服务发现。只有账号可见且开放对应模式的 Bernini 服务可以调用；公开别名来自当前服务，不能用教程中的占位符直接调用。

四种输入模式

模式	generationMode	输入素材
文生视频	`t2v`	不传素材
图像参考	`r2v`	1–5 张参考图
视频参考	`v2v`	一个源视频
图像加视频	`rv2v`	一张参考图和一个源视频

图片和视频分别按 素材上传 上传并保存 ID。当前输出以 480p/720p、横屏/竖屏、固定 5 秒为基础，最终允许组合仍以该服务能力为准。模式不同不意味着可以任意组合素材数量。

完整请求示例

以下四个请求各自独立；选择一个保存为 `request.json`：

```
{"model":"<MODEL_ALIAS>","generationMode":"t2v","prompt":"海面上的帆船，镜头缓慢拉远","resolutionTier":"480p","orientation":"landscape","seconds":5}
```

```
{"model":"<MODEL_ALIAS>","generationMode":"r2v","prompt":"保持参考图中主体的一致性，缓慢转动镜头","input_image_asset_ids":["<IMAGE_ASSET_ID>"],"resolutionTier":"480p","orientation":"landscape","seconds":5}
```

```
{"model":"<MODEL_ALIAS>","generationMode":"v2v","prompt":"保持源视频运镜，将场景转换为水彩风格","input_video_asset_id":"<VIDEO_ASSET_ID>","resolutionTier":"480p","orientation":"landscape","seconds":5}
```

```
{"model":"<MODEL_ALIAS>","generationMode":"rv2v","prompt":"结合参考图主体与源视频运镜","input_image_asset_ids":["<IMAGE_ASSET_ID>"],"input_video_asset_id":"<VIDEO_ASSET_ID>","resolutionTier":"480p","orientation":"landscape","seconds":5}
```

提交、查询、下载

```
curl --fail-with-body --max-time 60 "$JUSP_BASE_URL/media/generations" \
  -H "Authorization: Bearer $JUSP_API_KEY" \
  -H 'Content-Type: application/json' \
  -H 'Idempotency-Key: bernini-shot-001' \
  --data-binary @request.json --output create.json
```

随后按 任务查询 等待终态并下载输出。当前输出计费时长为 5 秒，不把输入素材总时长当作输出计费时长，也不自行增加模式倍数。

常见输入错误

- `t2v` 携带素材、`r2v` 超过 5 张图、`rv2v` 使用多图：修改到该模式的数量约束。
- 素材过期或不属于当前身份/模型：使用同一身份、目标模型重新上传，不引用其他模型的输入资产。
- 400 规格错误：重新查询服务，不猜测模型可能支持的分辨率或时长。
- 网络超时：保留原业务键和 `jobUrl`，先确认原任务，不立即重建。

/docs/models/video/upscale

视频超分

视频超分提升已有视频清晰度，不需要提示词。按 快速开始 配置环境变量，并从 服务发现 选择可用的视频超分模型。

1. 上传源视频

```
curl --fail-with-body --max-time 120 "$JUSP_BASE_URL/assets/input?model=$JUSP_MODEL" \
  -H "Authorization: Bearer $JUSP_API_KEY" \
  -F 'video=@source.mp4' --output upload.json
```

只接受一个 `video/mp4` 文件，通用上传大小上限为 128 MiB；以实时服务约束为准。保存 `asset.assetId`。不限视频总时长不等于不限上传大小、存储或运行资源。

2. 选择目标规格并创建

当前目标短边包括 480p、720p、1080p、1440p，选择当前服务实际开放的一项。输出保持原画幅；不传 `orientation`、`seconds` 或 `prompt`。平台选择最大的 `4n+1` 有效帧前缀，例如 150 帧按 149 帧处理，通常无需手工裁剪。

```
{"model":"<MODEL_ALIAS>","input_video_asset_id":"<VIDEO_ASSET_ID>","resolutionTier":"720p","seed":42}
```

将完整请求保存为 `request.json`：

```
curl --fail-with-body --max-time 60 "$JUSP_BASE_URL/media/generations" \
  -H "Authorization: Bearer $JUSP_API_KEY" \
  -H 'Content-Type: application/json' \
  -H 'Idempotency-Key: upscale-video-001' \
  --data-binary @request.json --output create.json
```

`batch_size=0` 和 `color_correction=adain` 属于当前默认处理配置；调用方不需要添加或自行调优内部执行参数。

3. 等待并读取结果

按 任务查询和下载 查询原 `jobUrl`，从 `outputs[].contentUrl` 下载 MP4，输出保留源视频音频。平台不因执行时间较长设置超分任务总时长截止；客户端仍需设置单请求超时和可配置的总等待预算。预算用完只结束本次等待，不取消服务端任务。

错误处理

非 MP4、无法解码、缺少有效帧信息或未开放的短边会被拒绝。429 等可恢复错误按服务端建议退避；任务失败保存 `jobId`、`requestId` 和错误码。不要用源视频时长乘一个固定倍数推算运行时间；价格以当前规格和有效帧计量为准。

X-Dub 人物口型同步

先完成 快速开始，确认 当前服务 向该 Key 开放口型同步能力。

准备两个素材

输入	multipart 字段	要求
人物视频	`video`	MP4，15–60 FPS，推荐 24–32 FPS，通用大小上限 128 MiB
驱动音频	`audio`	WAV，最多 60 秒，通用大小上限 25 MiB

当前处理会归一化到 25 FPS。使用清晰、稳定、可见的人脸和干净音频；输入质量会影响效果，但不能保证所有遮挡、侧脸都能成功。容量和格式若有变化，以当前服务的素材约束为准。

```
curl --fail-with-body --max-time 120 "$JUSP_BASE_URL/assets/input?model=$JUSP_MODEL" \
  -H "Authorization: Bearer $JUSP_API_KEY" -F 'video=@person.mp4' --output video.json
curl --fail-with-body --max-time 120 "$JUSP_BASE_URL/assets/input?model=$JUSP_MODEL" \
  -H "Authorization: Bearer $JUSP_API_KEY" -F 'audio=@speech.wav' --output audio.json
```

分别保存响应的 `asset.assetId`，不要互换音频和视频 ID。

创建请求

```
{"model":"<MODEL_ALIAS>","input_video_asset_id":"<VIDEO_ASSET_ID>","input_audio_asset_id":"<AUDIO_ASSET_ID>","seed":42}
```

保存为 `request.json` 后提交：

```
curl --fail-with-body --max-time 60 "$JUSP_BASE_URL/media/generations" \
  -H "Authorization: Bearer $JUSP_API_KEY" \
  -H 'Content-Type: application/json' \
  -H 'Idempotency-Key: lip-sync-001' \
  --data-binary @request.json --output create.json
```

该模式不需要提示词；不要传 `generationMode`、`resolutionTier`、`orientation`、`seconds`、`fps`、`enableRife` 或 `speedMultiplier`。它不是人物动作替换，不能改用 Wan 的 `characterMode`。

完成与计费

按 任务查询 查询原任务并下载输出。按驱动音频的有效时长、video_second 及当前生效单价结算；当前价格档使用 480p 不表示把输入画布强制缩为 480p。以输出 mimeType、实际视频和服务返回用量为准。

素材过期时重新上传；参数错误时先修正输入；HTTP 200 的任务查询仍可能是 failed，必须读取任务状态。保留原任务信息再决定是否重试，避免重复扣费。

/docs/models/audio

语音、音效与转写

语音合成

POST /audio/generations 创建异步音频任务。Qwen3-TTS 类服务使用 text 和受控音色字段，而不是所有场景都使用 prompt。

```
{"model":"<TTS_MODEL_ALIAS>","text":"欢迎使用星海智算。","voiceMode":"text"}
```

模式	字段	提醒
普通朗读	voiceMode=text	不上传视频、不猜预计音频秒数。
内置音色	voiceMode=custom_voice，speaker	speaker 使用模型支持的枚举，例如服务允许时的 serena。
音色描述	voiceMode=voice_design，instruct	只在服务开放此模式时使用。

可用音色见 音色目录 与 详情，以及目标服务的实际能力。音乐生成或其他音频模型按其请求合同选择参数，不复用 TTS 的字段猜测。

为视频生成音效

先以 video 字段上传受支持的视频，再向音效模型提交：

```
{"model":"<FOLEY_MODEL_ALIAS>","prompt":"与画面同步的脚步和衣物摩擦声","input_video_asset_id":"<VIDEO_ASSET_ID>"}
```

它生成音频，不自动替你把音轨合并回原视频。使用 任务查询和下载，按真实 MIME 保存输出。

语音转写

Qwen3-ASR-1.7B 把音频转换为文字。调用 POST /audio/transcriptions，在**本地文件**和**公网音频 URL**之间选择一种输入即可。

这是同步接口：一次请求直接返回转写结果，不返回 `jobId`，无需轮询或下载任务结果。模型广场的「API 接入」可以根据所选音频来源生成 Python、JavaScript 和 cURL 示例。

1. 准备 API Key 和模型别名

在门户的 API Key 管理页创建有权调用目标服务的 Key，并确认账户余额。Key 仅放在服务端环境变量中，不要写入前端页面或公开仓库。其他 ASR 服务请使用对应模型卡片的别名。

```
export JUSP_BASE_URL="https://api.jusuanhub.com/v1"
export JUSP_MODEL="qwen3-asr-1.7b"
export JUSP_API_KEY="<YOUR_API_KEY>"
```

2. 选择一种音频来源

本地文件：multipart 上传

将 `speech.wav` 替换为运行命令的机器上可读取的音频路径。cURL 会读取文件并上传；不要手动设置 multipart 的 `Content-Type`，也不要把 Python `bytes` 放进 JSON。

```
curl --fail-with-body --max-time 300 "$JUSP_BASE_URL/audio/transcriptions" \
  -H "Authorization: Bearer $JUSP_API_KEY" \
  -F "model=$JUSP_MODEL" \
  -F "file=@speech.wav" \
  -F "response_format=json"
```

公网音频 URL：JSON 请求

将示例地址替换为实际可下载的音频 HTTPS 地址。地址应直接返回音频文件，无需 Cookie 或登录；可使用有效期内的签名 URL，并给下载和转写留足有效期。网页分享链接、私网地址、本机路径不能作为音频 URL。

```
curl --fail-with-body --max-time 300 "$JUSP_BASE_URL/audio/transcriptions" \
  -H "Authorization: Bearer $JUSP_API_KEY" \
  -H "Content-Type: application/json" \
  --data "{\"model\":\"$JUSP_MODEL\",\"input_audio_url\":\"https://example.com/speech.wav\",\"response_format\":\"json\"}"
```

`file` 与 `input_audio_url` 必须且只能提供一项。URL 模式由平台下载音频，再校验实际格式、大小和时长；它不会绕过文件输入限制，也不需要先调用素材上传接口。

3. 读取转写结果

默认使用 JSON。下面是结构示例，文字和时长以实际音频为准；其他字段按目标服务能力返回。

```
{"text":"欢迎使用语音转写。","language":"Chinese","duration":3.2}
```

从 `text` 读取转写文字。Qwen3-ASR-1.7B 也支持 `response_format=text`，此时响应体是纯文本，不能调用 `response.json()`。

以下 Python 示例使用 URL 输入。先安装 `requests`，并沿用前面的环境变量：

```
import os
import requests

response = requests.post(
    f"{os.environ['JUSP_BASE_URL'].rstrip('/')}/audio/transcriptions",
    headers={"Authorization": f"Bearer {os.environ['JUSP_API_KEY']}"},
    json={
        "model": os.environ["JUSP_MODEL"],
        "input_audio_url": "https://example.com/speech.wav",
        "response_format": "json",
    },
    timeout=300,
)
response.raise_for_status()
print(response.json()["text"])
```

输入限制与可选参数

Qwen3-ASR-1.7B 的当前输入限制为：单个非空音频不超过 **25 MiB、300 秒**。常见格式包括 WAV、MP3、M4A、AAC、WebM、OGG 和 FLAC；文件内容必须与声明格式相符。其他模型以模型卡片和 服务能力 为准。

参数	用法
`model`	必填，目标服务的公开模型别名。
`file`	本地文件模式必填，通过 multipart 上传一个文件。
`input_audio_url`	URL 模式必填，通过 JSON 提交公网 HTTPS 音频地址。不能同时提供 `file`。
`language`	可选，省略时自动识别；仅使用目标模型支持的语言值。
`response_format`	Qwen3-ASR-1.7B 支持 `json`（默认）或 `text`。

语音转写按音频时长计量，计费秒数向上取整；单价以模型卡片当前价格为准。

调用失败时如何处理

- **400 / 422**：检查是否遗漏音频来源、同时提供两种来源、使用不支持的参数或格式；URL 是否能直接下载实际音频。
- **401 / 403**：检查 API Key 是否有效，以及是否有目标服务的调用权限。
- **402**：检查账户可用余额。
- **413**：音频或请求超过大小限制，请缩小输入。

- **429 / 5xx / 客户端超时**：保留响应中的请求标识并排查调用记录。同步转写没有任务轮询接口；结果未知时不要自动重复提交，避免重复计费。

更多鉴权与排查方式见 首次调用 和 错误处理。

/docs/models/prompt-optimization

提示词优化：H3 默认集成与 Qwen 手动改写

Qwen Image 2.1 手动改写

Qwen Image 2.1 使用同一 POST /media/prompt-optimizations 独立优化入口，由既有 Qwen3.8 按官方改写规则调用一次并返回正文。它为手动可选，不默认自动生成，不套用 H3 的结构化 JSON 或修复流程。

```
{
  "clientRequestId": "qwen-rewrite-001",
  "model": "qwen-image-2.1",
  "generationMode": "i2i",
  "originalPrompt": "将图1的人物放入图2的场景，保持人物特征。",
  "mediaSpec": {"outputSizeCode": "qwen21_standard_1376x768"},
  "referenceInputs": [
    {"role": "reference_image_1", "assetId": "<ASSET_1>"},
    {"role": "reference_image_2", "assetId": "<ASSET_2>"}
  ]
}
```

使用当前 Key 先上传参考图；请求头 `Idempotency-Key` 与 `clientRequestId` 相同。纯文字使用 `generationMode: t2i` 和空数组；单图或多图使用 `i2i` 和原序 `reference_image_N`。十张生成参考图合法，但优化模型有自己的上限，全部图片会按原序转交；上游拒绝时返回错误，不静默丢图或增加调用次数。

读取 `optimization.optimizedPrompt`，展示候选供用户编辑和采用；采用后将文本填入图片生成的 `prompt`，保留原 `size`，再以新的业务键提交生成。HTTP 结果仍是 JSON 封套，其中优化正文为普通文本；不要求模型正文输出 JSON。优化失败不自动生成，用户可选择保留原文。优化按 Qwen3.8 用量独立计费，后续生成另计费。

默认已包含在 H3 生成流程中

使用 Account API Key 调用 H3 的 `POST /v1/media/generations` 时，平台默认在同一个 Job 内先优化提示词，再生成视频。**只需提交一次生成请求，无需先额外调用优化接口。** 这适用于文生视频、首帧生视频、首尾帧生视频和全能参考四种模式。

用法	如何提交	是否调用优化
默认优化后生成	省略 `optimizePrompt`，或传 `true`	在生成 Job 内完成优化

用法	如何提交	是否调用优化
直接使用确定的文本	生成请求顶层传 `"optimizePrompt": false`	不调用优化、不产生优化费用
单独优化，查看或编辑后生成	先调用独立优化接口；采用结果生成时传 `false`	独立优化一次，生成时不再优化

`optimizePrompt` 是生成请求的顶层参数，不放进 `mediaSpec`；只接受 JSON 布尔值。关闭优化不会关闭内容安全审核或视频生成计费。平台应用调用 H3 也支持默认优化及同样的关闭参数；应用开发者参见内部应用开发中的平台服务教程。个人创作及其他模型不自动套用这个默认规则。

默认调用示例：

```
{"model":"minimax-h3","generationMode":"t2v","prompt":"清晨海岸，镜头缓慢推
进","resolutionTier":"544p","orientation":"landscape","seconds":5}
```

无需自动优化时，在同一正文添加 `"optimizePrompt": false`。完整的提交、查询和下载示例见 API Key 异步媒体调用。优化失败会终止任务，不回退原文继续生成；费用按实际优化调用结算。

平台负责保持参考素材的类型、顺序和编号。优化正文未逐一写出全部合法素材编号，不会仅因此判定失败或重复调用优化模型；实际素材缺失、引用不存在的编号、输出截断、超长以及安全审核错误仍按各自原因处理。不要把编号出现等同于模型完全理解了素材，复杂参考关系仍建议检查优化结果。

优化要求默认保留原提示词中对白的原文与语种；明确指定某句、某个角色或全局台词语种时，按对应范围忠实翻译。英文镜头描述不代表允许把中文台词改成英文。优化结果仍应核对原意，尤其是长对白与短视频时长互相冲突的情况。

需要单独优化时

独立优化适合先查看、编辑、采用或放弃候选的场景。POST /media/prompt-optimizations 只返回优化结果，不自动生成媒体；采用结果后再生成是另一个业务请求，使用独立幂等键。

先查询当前服务的能力与素材限制。`automaticPromptOptimization` 描述公开 H3 的自动编排是否启用；独立优化由 `promptOptimization` 声明。自动编排就绪不等于所有模型后端实时健康，能力不可用时须处理明确错误。

构造请求

```
{
  "clientRequestId":"your-optimization-business-id",
  "model":"<MODEL_ALIAS>",
  "generationMode":"t2v",
  "originalPrompt":"雨夜街道上的撑伞人物",
  "mediaSpec":{"resolutionTier":"768p","orientation":"landscape","seconds":5},
  "referenceInputs":[]
}
```

请求头 `Idempotency-Key` **必须等于** `clientRequestId`。图生、首尾帧和全能参考要附匹配的素材；优化接口的每个参考项只传 `role` 与 `assetId`，不要把生成请求的 `mediaType` 直接复制进去。

```
{"role":"reference_image","assetId":"<ASSET_ID>"}
```

音频参考按服务公开策略保留用途，不代表模型理解音频内容。有 H3 音频参考时须传 `audioReferencePolicy: "preserve_without_understanding"`。不要静默丢弃音频或其他素材。

优化时用于识图的图片会按原比例缩小至长边不超过 1024 像素，小图不放大、不裁切，透明区域以白底显示。视频生成仍使用原始素材；无需另行上传缩略图。细小文字或细节可能因缩小而不易辨认，关键要求应同时写在原提示词中。

确认再生成

1. 优化成功读取 `optimization.optimizedPrompt`。
2. 展示给用户，允许修改、采用或放弃；失败时保留原提示词。
3. 用户确认后，将最终文本放到生成请求的 `prompt`；公开 H3 同时传 `optimizePrompt: false`，避免再优化一次。
4. 使用**另一个生成业务幂等键**提交媒体请求，进入 任务查询。

例如，独立优化后采用候选的 H3 生成请求为：

```
{"model":"<MODEL_ALIAS>","generationMode":"t2v","prompt":"人工采用并确认的最终正
文","optimizePrompt":false,"resolutionTier":"768p","orientation":"landscape","seconds":5}
```

个人创作继续采用独立优化、人工采用、生成流程。App 的 H3 生成入口支持默认优化及顶层 `optimizePrompt`；采用独立优化结果生成时同样传 `false`，避免再次优化。SDK 接入方式见内部应用开发中的平台服务教程。

自动优化、失败与费用

公开 H3 `POST /v1/media/generations` 的 `optimizePrompt` 缺省为 `true`，显式 `false` 跳过优化和优化费用。只接受 JSON 布尔值；`null`、字符串或数字均无效。其他模型不因省略字段自动优化。关闭优化仍保留正常生成审核和生成费用。

平台尚未启用自动优化能力时，H3 缺省或 `optimizePrompt: true` 请求返回 `503 prompt_optimization_capability_unavailable`。此时不会按原文自动生成；需要直接使用已确认文本时显式传 `false`，仍需满足正常生成审核与规格要求。默认优化示例以平台已启用该能力为前提。

开启后创建响应继续是异步 Job，202 仅表示任务受理。轮询原 `jobUrl`；`promptOptimization` 可显示准备状态及尝试次数，不能将准备中当作视频已经开始。已知可重试技术失败最多重试一次；技术重试耗尽时 Job 为 `failed`，`errorCode` 为 `prompt_optimization_failed`，不回退原文生成。安全拒绝、需要澄清或候选拒绝不触发技术重试。

每个自动任务最多两次常规优化尝试，每次仅可对完整超长候选压缩一次，因此最多四次优化模型调用。常规调用、重试、压缩均按实际用量计费，已执行的调用即使任务失败也可能收费；未开始的视频生成不扣视频生成费用。独立生图优化最多一次常规调用加一次受控压缩，共两次模型调用，不增加 1200 字符硬上限，也不机械截断。

响应未知时先恢复原优化尝试；不能新建业务键盲目再调用或重复扣费。自动准备与生成共用任务创建起 30 分钟期限，重试不重新获得期限。重放同一请求复用原 Job 与已冻结策略，不重复优化；已受理的历史任务不会因默认值变化追加优化。

自动优化的错误会区分原因：准备超时为 prompt_optimization_timeout；内容审核未通过为 content_safety_blocked；需要补充说明为 content_safety_requires_clarification；优化稿被拒绝为 prompt_optimization_candidate_rejected；审核服务暂不可用为 content_safety_unavailable。这些错误发生于 failureStage：prompt_optimization，安全相关原因的 phase 为 content_safety。技术失败继续使用 prompt_optimization_failed。响应说明视频生成尚未开始，优化费用仍按实际调用结算。

查询 HTTP 200 仅表示查询成功，必须检查正文 status、errorCode、errorMessage；有 details 时可显示受控的审核原因和操作指导。retryable：false 表示不要自动新建付费任务。恢复查询保留原失败原因。

限流和未知结果

并发、频率和服务端超时读取当前能力合同，不在客户端写死套餐额度。429 或明确的容量保护按返回建议退避；结果未知时保留原正文和原键恢复。

optimization_in_progress 表示同一操作尚在处理，不应新建键再优化。明确终态失败后的新动作使用新业务键，并向用户说明；不能把失败直接转换为未经确认的视频生成。

/docs/models/assets

上传和读取素材

输入素材有两种提供方式：先上传得到素材 ID，或在创建请求里直接给公网 HTTPS 地址。两者按槽位二选一。

直接使用公网 URL

图像、视频、音频生成与文档 OCR、解析接口里，每个素材 ID 字段都有并列的 URL 字段：单个 *_asset_id 对应 *_url，数组 *_asset_ids 对应 *_urls；H3 全能参考的 referenceInputs[] 每行用 url 代替 assetId。

```
{"model":"<MODEL_ALIAS>","generationMode":"i2v","prompt":"镜头缓慢推
进","input_image_url":"https://example.com/first.png","resolutionTier":"768p","orientation":"landscape","seconds":5}
```

规则	内容
地址	只接受公网 HTTPS；图片槽位另接受 data:image/...
处理	平台下载并按与上传相同的格式、大小、时长与角色限制校验，作为本次请求的一小时临时素材，不进入“我的资源”
冲突	同一槽位同时给 ID 与 URL 返回 400 ambiguous_input_source
失败	invalid_source_url（地址不合规）、source_fetch_failed、source_fetch_timeout（下载失败或超时），修正地址或改用上传

同一请求重复出现的 URL 只下载一次。素材需要在多个任务间复用，或源站不允许平台访问时，仍先上传再引用 ID。

上传输入

POST /assets/input 使用 multipart，查询参数 `model` 必填。图片、视频和音频按服务合同选择 `image`、`video`、`audio` 字段；通用文件和文档使用 `file`。不要用 JSON 编码原始二进制。

```
curl --fail-with-body --max-time 120 \
  "$JUSP_BASE_URL/assets/input?model=$JUSP_MODEL" \
  -H "Authorization: Bearer $JUSP_API_KEY" -F 'image=@reference.png'
```

示例中的模型别名由平台提供；自己拼接查询参数时应 URL 编码。成功响应读取 `asset.assetId`，不是顶层 `assetId`。上传不等于启动模型生成。

保留素材边界

支持格式、字节大小、像素、时长和素材数量由目标模型及模式决定。提交前根据模型详情校验；不得通过改扩展名绕过格式限制。

文件大小与媒体时长是独立条件，不能按字节数推算秒数；压缩率和码率都会影响文件大小。H3 的各角色素材限制同时适用于 V1 上传后的任务与 V2 外部素材摄入。规格发现可读取 `GET /v1/services` 对应模型的 `mediaSpecContract.inputAssets.roleLimits`，按 `generationMode` 和 `role` 匹配。

H3 图片上传尚未指定首帧、尾帧或参考图角色时，按支持角色的通用安全上限做暂存准入；创建、估价和实际使用仍按所选角色重新校验。例如，长边6000像素且符合其他首帧限制的图片可以暂存并用作首帧，但超过参考图的5760像素上限，不能用作 `reference_image`。上传成功不代表适用于所有角色。

输入素材通常是临时资源，应读取响应的有效期，不把它当永久文件仓库。API 上传素材与 Portal“我的资源”上传列表不等价；不要假定任何可见资源 ID 都能跨模型、跨身份使用。

H3 全能参考中的音频使用 `audio` 字段；包含音轨的 MOV 作为音频参考时同样按音频身份上传，不消耗视频参考名额。上传身份不保证任意编码都可解码。

上传失败时保留具体原因

上传接口返回非2xx时，先保存原始 HTTP 状态、响应 JSON 的 `error.code`、`error.message`、`error.requestId`（或顶层 `requestId`）及 `X-Request-ID`。不要只记录“上传失败”或把上游400统一转成502后丢弃原始原因。V2 创建请求使用自己的错误格式：保留 `error.code`、`error.message` 和顶层 `request_id`。

400通常表示素材或请求未满足合同，例如视频时长越界、宽高比越界、格式与实际内容不符或无法解析；413表示请求或文件过大；415表示媒体格式不受支持。以本次响应的具体原因判断，修正素材后再提交，不对同一无效素材循环重试。文件很小也可能超过时长上限，改扩展名或 `Content-Type` 不能修复实际容器或编码问题。

向平台排查时提供请求时间、模型、原始状态、错误码、错误消息和请求追踪号；可附已移除 Authorization 的请求结构及不含私人内容的素材元信息。不要提供 API Key、签名 URL 或原始私人素材。若响应正文没有保留，请先提供请求追踪号，平台会依据可用日志核查；不能仅凭HTTP状态或响应字节数确定具体原因。

下载生成结果

任务成功后优先读取 `outputs[].contentUrl`，携带创建任务的同一 API Key。原样 URL 默认返回无可见水印的 source；只有显式追加 `deliveryMode=marked` 才生成带平台标识版本，并可能先返回 202。下载前可用对应 HEAD 操作检查元信息。不要拼接对象存储地址，不要将带鉴权的信息转发到不可信 Host。

生产客户端应校验响应 URL 与可信 API 基础地址同源，限制重定向，设置读超时和最大文件预算。示例中的链接不是匿名永久分享地址。

异步任务与下载 · GET 素材内容。

/docs/models/jobs

等待任务并下载结果

创建只表示受理

媒体创建接口成功通常返回 HTTP 202、`jobId`、绝对 `jobUrl`。`Location` 指向查询地址，`Retry-After` 或正文中的轮询建议控制查询节奏。

保存原任务 ID、查询地址和本次业务幂等键。支持 `Idempotency-Key` 的操作使用稳定业务键；同一业务动作的重试必须保持身份、正文和键一致。开始新业务动作才使用新键。

> `jobUrl` 和成功响应中的 `outputs[].contentUrl` 都是受保护地址。每次 GET 仍须携带创建任务时相同的 `Authorization: Bearer ...`。未鉴权的资源请求可能返回隐私型 404；不能据此判断资产不存在，也不能把该 URL 直接放入浏览器媒体标签。完整示例见 API Key 异步媒体完整调用。

按状态推进

状态	客户端处理
`queued`	继续等待，显示排队状态，不承诺完成时间。
`running`	显示执行状态；服务返回阶段信息时使用 `stage` / `stageMessage`。
`succeeded`	读取 `outputs[]`，原样 `contentUrl` 默认下载无可见水印的 source；需要带标识版本时追加 `deliveryMode=marked` 并处理可能的 202。
`failed`	停止轮询，读取错误和可重试标志；不自动再次创建。
历史 `canceled`	停止轮询；公开 API 当前不提供用户取消任务操作。

查询 HTTP 200 也可能是失败任务。Account API Key 当前没有独立的媒体 Job SSE 端点，不能使用应用 Runtime 的事件地址替代。

有界轮询示例（Python 标准库）

把创建响应保存为 `create.json`。此脚本只查询原任务，不创建新任务；原样下载第一项无可见水印的 source 输出到 `result.bin`。真实格式以响应 `mimeType` 为准。

```
import json
import os
import time
import urllib.parse
import urllib.request

base = os.environ["JUSP_BASE_URL"].rstrip("/")
origin = urllib.parse.urlsplit(base)
key = os.environ["JUSP_API_KEY"]
with open("create.json", encoding="utf-8") as file:
    created = json.load(file)

class NoRedirect(urllib.request.HTTPRedirectHandler):
    def redirect_request(self, req, fp, code, msg, headers, newurl):
        return None

opener = urllib.request.build_opener(NoRedirect)

def read(url):
    target = urllib.parse.urlsplit(url)
    if (target.scheme, target.netloc) != (origin.scheme, origin.netloc):
        raise ValueError("拒绝向非 API 来源发送凭据")
    request = urllib.request.Request(url, headers={"Authorization": "Bearer " + key})
    return opener.open(request, timeout=60)

deadline = time.monotonic() + 1800
while time.monotonic() < deadline:
    with read(created["jobUrl"]) as response:
        job = json.load(response)
    if job["status"] == "succeeded":
        if not job.get("outputs"):
            raise RuntimeError("任务成功但未返回输出，请保留 jobId 联系支持")
        output = job["outputs"][0]
        with read(output["contentUrl"]) as response, open("result.bin", "wb") as file:
            total = 0
            while chunk := response.read(1024 * 1024):
                total += len(chunk)
                if total > 1024 * 1024 * 1024:
                    raise ValueError("输出超过本示例 1 GiB 下载预算")
                file.write(chunk)
        break
    if job["status"] in {"failed", "canceled"}:
        raise RuntimeError(f"任务结束：{job.get('errorCode', job['status'])}")
    if job["status"] not in {"queued", "running"}:
        raise RuntimeError(f"未知任务状态：{job['status']}")
    delay = max(1, min(60, int(job.get("retryAfterSeconds") or job.get("refreshAfterSeconds") or 3)))
    time.sleep(min(delay, max(0, deadline - time.monotonic())))
else:
    raise TimeoutError("本地等待结束；保留原 jobUrl，稍后继续查询")
```

示例遇到 HTTP 错误会停止。生产客户端可对明确可恢复的读取失败按 错误恢复 退避，但不得因此重新提交生成请求。达到本地期限不表示服务端任务已经取消。

只有调用方显式在 `contentUrl` 后追加 `deliveryMode=marked` 时，下载接口才会进入带标识导出，并可能返回 `202`、`Retry-After` 和 `retryAfterSeconds`。此时应重试同一下载 URL，不能重新创建媒体 Job，也不能把 202 JSON 保存为结果文件。

/docs/models/errors

处理错误、限流和未知结果

先识别失败发生在哪一步

连接建立、HTTP 拒绝、任务业务失败和输出下载失败有不同恢复方式。保存安全的 `requestId`、状态码和机器错误码用于排查；不记录 Key、用户完整输入和带身份信息的 URL。

现象	应做什么	不应做什么
400 / 422	按字段错误修正模式、素材或规格	原样高速重试
401	检查 Key 完整性、失效或到期	在日志打印密钥确认
402	检查额度与当前价格准入	更换用户身份规避计费
403	检查服务范围、冻结和权限	自动扩大 scope
404	核对原模型、任务或素材 ID	新建任务“找回”旧结果
409	区分幂等冲突、处理中与资源冲突，查询原状态	用同键发送不同正文
429	遵守 `Retry-After`，限制并发并退避；`ACCOUNT_CONCURRENCY_LIMITED` 表示账号同时进行的同步请求已满，等已有请求完成再发	用多个 Key 绕过限制（同一账号的 Key 共用额度）
503 / 504	判断请求是否已经受理；有任务 ID 就继续查询原任务	无条件换键重提
查询返回 failed	停止轮询，读取业务错误；明确发起新动作才重试生成	把 HTTP 200 当作生成成功
SSE 提前断开	标记部分内容、保留用户输入	把截断内容作为完整结果

OpenAI SDK 错误对象

公开模型接口返回非 2xx 时，官方 OpenAI SDK 会抛出异常。平台同时返回旧客户端可读的顶层字段和标准兼容的嵌套 `error`：

```
{
  "error": {
    "message": "max_tokens maximum is 4096",
    "type": "invalid_request_error",
    "param": "max_tokens",
    "code": "invalid_argument",
    "requestId": "req_example",
    "retryable": false,
    "details": {"limit": 4096}
  },
  "code": "invalid_argument",
  "message": "max_tokens maximum is 4096",
  "requestId": "req_example",
  "retryable": false
}
```

Python SDK 常用字段是 `APIStatusError.status_code`、`code`、`param`、`request_id` 和 `body`；Node/TypeScript SDK 常用字段是 `APIError.status`、`code`、`param`、`requestID` 和 `error`。完整正文在 `body/error` 中，`details` 用于限制值等机器信息。应用应保存状态码、错误码和请求 ID，不要依赖自然语言正文解析。

`retryable=true` 才表示平台允许按该错误的恢复语义重试；429 还应遵守 `Retry-After`。请求已经发往模型但结果未知时，即使是 5xx 也不能盲目重放；若接口支持幂等，应使用同一身份、同一业务键和同一正文查询或恢复。

流式响应如果已经产生部分内容后发生断流，SDK 会收到流错误；部分文本必须标记为不完整，不得当作成功结果或补写伪造的 `[DONE]`。

写入结果未知

响应丢失不等于服务端没有执行。公开操作声明支持幂等时，用**同身份、同业务键、同正文**恢复；已经取得 ID 时优先查询。未声明幂等的操作不自动重放。

API Key 创建不提供秘密找回，遵循 创建结果不确定。一次性会话交换与模型生成不是同一种重试协议。

计费与等待

请求可能先进行额度预占，再按最终执行事实结算。页面报价不是永久锁定价格；最终由真实请求命中的规则决定。上传成功、任务受理和模型完成是三个不同事实。

缩短客户端超时不会加速执行，也不会自动取消已受理任务。为每条链设置总等待预算和最大退避时间；不要对网络问题无限轮询。

预检媒体准入

POST /v1/admission/preview

身份与入口

使用公开模型 API 基础地址和 Account API Key；不要混用应用 RuntimeSession。具体模型的条件字段由实时服务合同声明。

- 基础地址：/

认证：[{"bearerAuth":[]}]；凭据说明见类型附录。

请求参数

位置	参数	必填	类型	说明
请求头	X-Client-Request-Id	否	string	可选客户端关联值，只能包含可见 ASCII，最多 512 字节；不能替代服务端 requestId。

请求正文

请求正文必填。

application/json

类型：AdmissionPreviewRequest。完整字段见本文末尾的类型定义附录。

字段	类型	必填	说明
model	string	是	支持异步准入的服务公开模型别名。
request	object / null	否	待预检的创建请求体；省略或为 null 时视为空对象。平台使用所选别名覆盖 request.model。

```
{
  "model": "minimax-h3",
  "request": {
    "generationMode": "t2v",
    "model": "minimax-h3",
    "orientation": "landscape",
    "prompt": "雨夜站台，镜头缓慢推进，人物自然递伞后共同离开。",
    "resolutionTier": "768p",
    "seconds": 5,
    "seed": 1234,
    "storageMode": "personal_30d"
  }
}
```

响应

HTTP 状态	说明	媒体类型与结构
200	请求成功	application/json：AdmissionPreviewResponse
400	请求参数不合法	application/json：PublicError
401	未认证或凭据失效	application/json：PublicError
402	额度不足	application/json：PublicError
403	没有访问权限	application/json：PublicError
404	资源不存在或不可见	application/json：PublicError
409	请求或状态冲突	application/json：PublicError
413	请求体超过大小限制	application/json：PublicError
415	不支持的媒体类型	application/json：PublicError
422	输入无法处理	application/json：PublicError
429	请求过于频繁	application/json：PublicError
500	服务内部错误	application/json：PublicError
502	上游响应异常	application/json：PublicError
503	服务暂不可用	application/json：PublicError
504	上游响应超时	application/json：PublicError

上传输入素材

POST /v1/assets/input

身份与入口

使用公开模型 API 基础地址和 Account API Key；不要混用应用 RuntimeSession。具体模型的条件字段由实时服务合同声明。

- 基础地址：/

认证：[{"bearerAuth":[]}]；凭据说明见类型附录。

请求参数

位置	参数	必填	类型	说明
查询	model	是	string	限定任务或素材所属模型范围的公开别名；须在当前 API Key 授权范围内。
请求头	X-Client-Request-Id	否	string	可选客户端关联值，只能包含可见 ASCII，最多 512 字节；不能替代服务端 requestId。

请求正文

请求正文必填。

multipart/form-data

类型：object / object / object / object。完整字段见本文末尾的类型定义附录。

每次上传一个文件，遵守 GET /v1/services 的 mediaSpecContract。通用上限：PNG/JPEG/WebP 图片 16 MiB，WAV 音频 25 MiB，MP4 视频 128 MiB。H3 全能参考：图片 30 MiB，MP4 视频 50 MiB，音频 15 MiB；支持 WAV/MP3/FLAC/M4A/AAC/OGG 或带音轨 MOV。MOV 须用 audio 字段上传，保留音频身份和原字节，不作为视频参考。H3 音频每段 2 ~ 15 秒、最多 3 段、合计不超过 15 秒；首尾帧图片仍使用各自 16 MiB 限制。

字段	类型	必填	说明
audio	string	否	上传音频文件的 multipart 字段。
file	string	否	上传原始文件的 multipart 字段，不使用 JSON 编码二进制。
image	string	否	上传图片文件的 multipart 字段。
storageMode	string	否	平台存储策略，使用顶层字段：默认 personal_30d，将本次主动上传或生成的资源保存到当前用户资源空间，自资源入库起 30 天到期，可在门户“我的资源”改为永久保存，仍占用配额。temporary_1h 自入库起 1 小时到期，仅后台已对所属用户或组织开启临时存储权限时可选；未授权时拒绝，不静默降级。生成请求不改变已引用素材的原到期时间。公网 URL 受控导入的临时输入遵循独立的 1 小时策略，不因本字段进入个人资源空间。；可选值：["personal_30d","temporary_1h"]；默认值："personal_30d"
video	string	否	上传视频文件的 multipart 字段。

响应

HTTP 状态	说明	媒体类型与结构
201	已创建输入素材，或在相同 API Key 和服务范围内复用已有素材	application/json：InputAssetUploadResponse
400	请求参数不合法	application/json：PublicError
401	未认证或凭据失效	application/json：PublicError
402	额度不足	application/json：PublicError
403	没有访问权限	application/json：PublicError
404	资源不存在或不可见	application/json：PublicError

HTTP 状态	说明	媒体类型与结构
409	请求或状态冲突	application/json：PublicError
413	请求体超过大小限制	application/json：PublicError
415	不支持的媒体类型	application/json：PublicError
422	输入无法处理	application/json：PublicError
429	请求过于频繁	application/json：PublicError
500	服务内部错误	application/json：PublicError
502	上游响应异常	application/json：PublicError
503	服务暂不可用	application/json：PublicError
504	上游响应超时	application/json：PublicError

/docs/reference/models/readInputAsset

读取输入素材

读取已签发的短期图片/视频 URL；仅有 Bearer Key 不足以授权。不要自行构造令牌、修改参数、记录 URL 或将本接口用作结果下载。令牌无效/过期返回 401 invalid_input_asset_token。

`GET /v1/assets/input/{id}/content`

身份与入口

使用公开模型 API 基础地址和 Account API Key；不要混用应用 RuntimeSession。具体模型的条件字段由实时服务合同声明。

- 基础地址：`/`

认证：[{"inputAssetReadToken":[]}]；凭据说明见类型附录。

请求参数

位置	参数	必填	类型	说明
路径	id	是	string	

位置	参数	必填	类型	说明
请求头	X-Client-Request-Id	否	string	可选客户端关联值，只能包含可见 ASCII，最多 512 字节；不能替代服务端 requestId。

响应

HTTP 状态	说明	媒体类型与结构
200	鉴权后的二进制媒体正文；Account API Key 结果默认返回 source，显式 deliveryMode=marked 返回带标识导出。HEAD 仅返回响应头，无 JSON 包装。读取失败可重试原读取，不重新创建任务。	application/octet-stream：string
200 响应头	Content-Length	
200 响应头	Content-Type	
200 响应头	ETag	
200 响应头	X-Jusp-Asset-Variant	实际返回的结果变体：source 或 marked_export。
400	请求参数不合法	application/json：PublicError
401	未认证或凭据失效	application/json：PublicError
402	额度不足	application/json：PublicError
403	没有访问权限	application/json：PublicError
404	资源不存在或不可见	application/json：PublicError
409	请求或状态冲突	application/json：PublicError
413	请求体超过大小限制	application/json：PublicError
415	不支持的媒体类型	application/json：PublicError
422	输入无法处理	application/json：PublicError
429	请求过于频繁	application/json：PublicError
500	服务内部错误	application/json：PublicError
502	上游响应异常	application/json：PublicError
503	服务暂不可用	application/json：PublicError

HTTP 状态	说明	媒体类型与结构
504	上游响应超时	application/json：PublicError

/docs/reference/models/readAsset

读取生成结果

GET /v1/assets/{id}/content

身份与入口

使用公开模型 API 基础地址和 Account API Key；不要混用应用 RuntimeSession。具体模型的条件字段由实时服务合同声明。

- 基础地址：/

认证：[{"bearerAuth":[]}]；凭据说明见类型附录。

请求参数

位置	参数	必填	类型	说明
路径	id	是	string	
查询	model	是	string	限定任务或素材所属模型范围的公开别名；须在当前 API Key 授权范围内。
查询	deliveryMode	否	string	结果交付模式。省略或 source 返回无可见水印的源文件；显式 marked 创建或复用带平台标识的导出，未就绪时返回 202。
请求头	X-Client-Request-Id	否	string	可选客户端关联值，只能包含可见 ASCII，最多 512 字节；不能替代服务端 requestId。

响应

HTTP 状态	说明	媒体类型与结构
200	鉴权后的二进制媒体正文；Account API Key 结果默认返回 source，显式 deliveryMode=marked 返回带标识导出。HEAD 仅返回响应头，无 JSON 包装。读取失败可重试原读取，不重新创建任务。	application/octet-stream：string
200 响应头	Content-Length	
200 响应头	Content-Type	
200 响应头	ETag	
200 响应头	X-Jusp-Asset-Variant	实际返回的结果变体：source 或 marked_export。
202	仅显式 deliveryMode=marked 时可能出现：带标识导出尚未就绪，按 Retry-After 重试同一 URL。	application/json：object
202 响应头	Retry-After	建议重试等待秒数。
400	请求参数不合法	application/json：PublicError
401	未认证或凭据失效	application/json：PublicError
402	额度不足	application/json：PublicError
403	没有访问权限	application/json：PublicError
404	资源不存在或不可见	application/json：PublicError
409	请求或状态冲突	application/json：PublicError
413	请求体超过大小限制	application/json：PublicError
415	不支持的媒体类型	application/json：PublicError
422	输入无法处理	application/json：PublicError
429	请求过于频繁	application/json：PublicError
500	服务内部错误	application/json：PublicError
502	上游响应异常	application/json：PublicError
503	服务暂不可用	application/json：PublicError
504	上游响应超时	application/json：PublicError

读取生成结果元信息

HEAD /v1/assets/{id}/content

身份与入口

使用公开模型 API 基础地址和 Account API Key；不要混用应用 RuntimeSession。具体模型的条件字段由实时服务合同声明。

- 基础地址：/

认证：[{"bearerAuth":[]}]；凭据说明见类型附录。

请求参数

位置	参数	必填	类型	说明
路径	id	是	string	
查询	model	是	string	限定任务或素材所属模型范围的公开别名；须在当前 API Key 授权范围内。
查询	deliveryMode	否	string	结果交付模式。省略或 source 返回无可见水印的源文件；显式 marked 创建或复用带平台标识的导出，未就绪时返回 202。
请求头	X-Client-Request-Id	否	string	可选客户端关联值，只能包含可见 ASCII，最多 512 字节；不能替代服务端 requestId。

响应

HTTP 状态	说明	媒体类型与结构
200	鉴权后的二进制媒体正文；Account API Key 结果默认返回 source，显式 deliveryMode=marked 返回带标识导出。HEAD 仅返回响应头，无 JSON 包装。读取失败可重试原读取，不重新创建任务。	
200 响应头	Content-Length	
200 响应头	Content-Type	
200 响应头	ETag	
200 响应头	X-Jusp-Asset-Variant	实际返回的结果变体：source 或 marked_export。
202	仅显式 deliveryMode=marked 时可能出现：带标识导出尚未就绪，按 Retry-After 重试同一 URL。	
202 响应头	Retry-After	建议重试等待秒数。
400	请求参数不合法	application/json：PublicError
401	未认证或凭据失效	application/json：PublicError
402	额度不足	application/json：PublicError
403	没有访问权限	application/json：PublicError
404	资源不存在或不可见	application/json：PublicError
409	请求或状态冲突	application/json：PublicError
413	请求体超过大小限制	application/json：PublicError
415	不支持的媒体类型	application/json：PublicError
422	输入无法处理	application/json：PublicError
429	请求过于频繁	application/json：PublicError
500	服务内部错误	application/json：PublicError
502	上游响应异常	application/json：PublicError
503	服务暂不可用	application/json：PublicError
504	上游响应超时	application/json：PublicError

生成音频

生成产物默认以 personal_30d 入库当前用户资源空间，占用空间配额，自入库起 30 天到期；可在门户改为永久保存。顶层 storageMode=temporary_1h 仅后台已对用户或组织授权时允许，自入库起 1 小时到期。引用既有资源不会改变其期限；公网 URL 临时输入继续按独立 1 小时策略处理。普通 API Key 不接受应用来源或隐藏产物字段。

`POST /v1/audio/generations`

身份与入口

使用公开模型 API 基础地址和 Account API Key；不要混用应用 RuntimeSession。具体模型的条件字段由实时服务合同声明。

- 基础地址：`/`

认证：[{"bearerAuth":[]}]；凭据说明见类型附录。

请求参数

位置	参数	必填	类型	说明
请求头	X-Client-Request-Id	否	string	可选客户端关联值，只能包含可见 ASCII，最多 512 字节；不能替代服务端 requestId。

请求正文

请求正文必填。

application/json

类型：`AudioGenerationRequest`。完整字段见本文末尾的类型定义附录。

字段	类型	必填	说明
generationMode	string	否	所选服务支持的生成模式。

字段	类型	必填	说明
input_video_asset_id	string	否	已上传的视频素材，用于视频音效；纯文本语音合成不传此字段。
input_video_url	string	否	可选：与 input_video_asset_id 并列的外部来源，二选一。只接受公网 HTTPS（443 端口）；平台在接纳前经受控下载器摄入为当前调用身份和模型的一小时临时素材，再按资源 ID 原链路处理，等同先调用 /v1/assets/input。同槽位同时传 ID 与 URL 返回 400 ambiguous_input_source；同请求重复 URL 只下载一次；来源被拒为 400 invalid_source_url，下载失败 502 source_fetch_failed，超时 504 source_fetch_timeout；部署关闭本能力时返回 400 input_url_not_enabled。
instruct	string	否	voice_design 模式必填；Qwen3-TTS 最多 160 个字符。
model	string	是	音频服务公开模型别名；语音、音乐与视频音效使用不同的条件字段。
prompt	string	否	服务支持时填写音效或音乐描述。
seconds	integer	否	请求时长，单位秒；仅适用于声明支持该字段的模型。
seed	integer	否	可选随机种子；幂等重试时保持原值。
speaker	string	否	custom_voice 模式必填；使用服务已公布的音色名称。

字段	类型	必填	说明
storageMode	string	否	平台存储策略，使用顶层字段：默认 personal_30d，将本次主动上传或生成的资源保存到当前用户资源空间，自资源入库起 30 天到期，可在门户“我的资源”改为永久保存，仍占用配额。temporary_1h 自入库起 1 小时到期，仅后台已对所属用户或组织开启临时存储权限时可选；未授权时拒绝，不静默降级。生成请求不改变已引用素材的原到期时间。公网 URL 受控导入的临时输入遵循独立的 1 小时策略，不因本字段进入个人资源空间。；可选值：["personal_30d","temporary_1h"]；默认值："personal_30d"
text	string	否	待合成的文本；Qwen3-TTS 的字符上限由服务声明，其他模型使用各自限制。；最长长度：600
voiceMode	string	否	Qwen3-TTS 模式：text（默认普通朗读）、custom_voice（内置音色）、voice_design（音色描述）。

响应

HTTP 状态	说明	媒体类型与结构
200	返回重放的异步任务，或同步执行服务的结果。H3 使用 MediaJob 分支；请核对所选服务的执行方式。	application/json：MediaJob / SynchronousMediaResponse
202	异步任务已受理或仍在执行；使用原 API Key 请求绝对 jobUrl（也见 Location），遵守 Retry-After 的查询间隔。	application/json：MediaJob
202 响应头	Location	带鉴权的绝对轮询地址，与正文 jobUrl 相同。
202 响应头	Retry-After	建议距离下次轮询的等待秒数。
400	请求参数不合法	application/json：PublicError
401	未认证或凭据失效	application/json：PublicError

HTTP 状态	说明	媒体类型与结构
402	额度不足	application/json：PublicError
403	没有访问权限	application/json：PublicError
404	资源不存在或不可见	application/json：PublicError
409	请求或状态冲突	application/json：PublicError
413	请求体超过大小限制	application/json：PublicError
415	不支持的媒体类型	application/json：PublicError
422	输入无法处理	application/json：PublicError
429	请求过于频繁	application/json：PublicError
500	服务内部错误	application/json：PublicError
502	上游响应异常	application/json：PublicError
503	服务暂不可用	application/json：PublicError
504	上游响应超时	application/json：PublicError

202 · application/json 示例


```
{
  "id": "job_example",
  "jobId": "job_example",
  "jobUrl": "https://api.example.test/v1/jobs/job_example?model=minimax-h3",
  "model": "minimax-h3",
  "requestId": "req_example",
  "serviceKind": "video_generation",
  "stage": "queued",
  "status": "queued"
}
```


/docs/reference/models/createAudioTranscription

转写音频

同步语音转写，直接返回结果，无需轮询任务。multipart/form-data 上传 file，或 application/json 提交 input_audio_url；两种输入互斥。音频大小、时长、语言和返回格式由所选服务声明。

POST /v1/audio/transcriptions

身份与入口

使用公开模型 API 基础地址和 Account API Key；不要混用应用 RuntimeSession。具体模型的条件字段由实时服务合同声明。

- 基础地址：`/`

认证：[{"bearerAuth":[]}]；凭据说明见类型附录。

请求参数

位置	参数	必填	类型	说明
请求头	X-Client-Request-Id	否	string	可选客户端关联值，只能包含可见 ASCII，最多 512 字节；不能替代服务端 requestId。

请求正文

请求正文必填。

application/json

类型：`AudioTranscriptionURLRequest`。完整字段见本文末尾的类型定义附录。

字段	类型	必填	说明
input_audio_url	string	是	无需登录即可下载的公网 HTTPS 音频地址，可使用有效期内的签名 URL；与 file 互斥。平台受限下载并验证实际格式、大小和时长。；匹配规则：^https://；格式：uri
language	string	否	仅在所选模型明确支持时传入。
model	string	是	语音转写服务的公开模型别名。
response_format	string	否	仅在所选模型支持时使用；通常返回 JSON。

multipart/form-data

类型：`AudioTranscriptionRequest`。完整字段见本文末尾的类型定义附录。

字段	类型	必填	说明
file	string	是	一个非空文件，不超过 25 MiB。MIME 支持 audio/wav、x-wav、mpeg、mp3、mp4、x-m4a、webm、ogg、flac；octet-stream 须有受支持的文件扩展名。；格式：binary
language	string	否	仅在所选模型明确支持时传入。
model	string	是	语音转写服务的公开模型别名。
response_format	string	否	仅在所选模型支持时使用；通常返回 JSON。

响应

HTTP 状态	说明	媒体类型与结构
200	请求成功	application/json：AudioTranscriptionResponse；text/plain：string
400	请求参数不合法	application/json：PublicError
401	未认证或凭据失效	application/json：PublicError
402	额度不足	application/json：PublicError
403	没有访问权限	application/json：PublicError
404	资源不存在或不可见	application/json：PublicError
409	请求或状态冲突	application/json：PublicError
413	请求体超过大小限制	application/json：PublicError
415	不支持的媒体类型	application/json：PublicError
422	输入无法处理	application/json：PublicError
429	请求过于频繁	application/json：PublicError
500	服务内部错误	application/json：PublicError
502	上游响应异常	application/json：PublicError
503	服务暂不可用	application/json：PublicError

HTTP 状态	说明	媒体类型与结构
504	上游响应超时	application/json：PublicError

/docs/reference/models/listVoices

查询公共音色列表

GET /v1/audio/voices

身份与入口

使用公开模型 API 基础地址和 Account API Key；不要混用应用 RuntimeSession。具体模型的条件字段由实时服务合同声明。

- 基础地址：/

认证：[{"bearerAuth":[]}]；凭据说明见类型附录。

请求参数

位置	参数	必填	类型	说明
请求头	X-Client-Request-Id	否	string	可选客户端关联值，只能包含可见 ASCII，最多 512 字节；不能替代服务端 requestId。

响应

HTTP 状态	说明	媒体类型与结构
200	请求成功	application/json：CatalogVoiceList
400	请求参数不合法	application/json：PublicError
401	未认证或凭据失效	application/json：PublicError
402	额度不足	application/json：PublicError
403	没有访问权限	application/json：PublicError

HTTP 状态	说明	媒体类型与结构
404	资源不存在或不可见	application/json：PublicError
409	请求或状态冲突	application/json：PublicError
413	请求体超过大小限制	application/json：PublicError
415	不支持的媒体类型	application/json：PublicError
422	输入无法处理	application/json：PublicError
429	请求过于频繁	application/json：PublicError
500	服务内部错误	application/json：PublicError
502	上游响应异常	application/json：PublicError
503	服务暂不可用	application/json：PublicError
504	上游响应超时	application/json：PublicError

/docs/reference/models/getVoice

查询公共音色详情

GET /v1/audio/voices/{id}

身份与入口

使用公开模型 API 基础地址和 Account API Key；不要混用应用 RuntimeSession。具体模型的条件字段由实时服务合同声明。

- 基础地址：/

认证：[{"bearerAuth":[]}]；凭据说明见类型附录。

请求参数

位置	参数	必填	类型	说明
路径	id	是	string	

位置	参数	必填	类型	说明
请求头	X-Client-Request-Id	否	string	可选客户端关联值，只能包含可见 ASCII，最多 512 字节；不能替代服务端 requestId。

响应

HTTP 状态	说明	媒体类型与结构
200	请求成功	application/json：CatalogVoice
400	请求参数不合法	application/json：PublicError
401	未认证或凭据失效	application/json：PublicError
402	额度不足	application/json：PublicError
403	没有访问权限	application/json：PublicError
404	资源不存在或不可见	application/json：PublicError
409	请求或状态冲突	application/json：PublicError
413	请求体超过大小限制	application/json：PublicError
415	不支持的媒体类型	application/json：PublicError
422	输入无法处理	application/json：PublicError
429	请求过于频繁	application/json：PublicError
500	服务内部错误	application/json：PublicError
502	上游响应异常	application/json：PublicError
503	服务暂不可用	application/json：PublicError
504	上游响应超时	application/json：PublicError

查询主账号积分余额

仅组织当前 active 主账号的 Account API Key 可调用：非组织成员返回 403 organization_membership_required，非主账号返回 403 organization_primary_required，组织停用返回 403 organization_access_suspended，组织上下文不可用返回 503 organization_context_unavailable。主账号身份每次请求实时校验。

`GET /v1/billing/balance`

身份与入口

使用公开模型 API 基础地址和 Account API Key；不要混用应用 RuntimeSession。具体模型的条件字段由实时服务合同声明。

- 基础地址：`/`

认证：[{"bearerAuth":[]}]；凭据说明见类型附录。

请求参数

位置	参数	必填	类型	说明
请求头	X-Client-Request-Id	否	string	可选客户端关联值，只能包含可见 ASCII，最多 512 字节；不能替代服务端 requestId。

响应

HTTP 状态	说明	媒体类型与结构
200	请求成功	application/json：BillingBalance
400	请求参数不合法	application/json：PublicError
401	未认证或凭据失效	application/json：PublicError
402	额度不足	application/json：PublicError
403	没有访问权限	application/json：PublicError

HTTP 状态	说明	媒体类型与结构
404	资源不存在或不可见	application/json：PublicError
409	请求或状态冲突	application/json：PublicError
413	请求体超过大小限制	application/json：PublicError
415	不支持的媒体类型	application/json：PublicError
422	输入无法处理	application/json：PublicError
429	请求过于频繁	application/json：PublicError
500	服务内部错误	application/json：PublicError
502	上游响应异常	application/json：PublicError
503	服务暂不可用	application/json：PublicError
504	上游响应超时	application/json：PublicError

/docs/reference/models/listBillingCalls

按变化游标查询调用明细

仅组织当前 active 主账号的 Account API Key 可调用：非组织成员返回 403 organization_membership_required，非主账号返回 403 organization_primary_required，组织停用返回 403 organization_access_suspended，组织上下文不可用返回 503 organization_context_unavailable。主账号身份每次请求实时校验。

GET /v1/billing/calls

身份与入口

使用公开模型 API 基础地址和 Account API Key；不要混用应用 RuntimeSession。具体模型的条件字段由实时服务合同声明。

- 基础地址：/

认证：[{"bearerAuth":[]}]；凭据说明见类型附录。

请求参数

位置	参数	必填	类型	说明
请求头	X-Client-Request-Id	否	string	可选客户端关联值，只能包含可见 ASCII，最多 512 字节；不能替代服务端 requestId。
查询	cursor	否	string	上一页返回的 nextCursor；与 since 同时提供时以 cursor 为准。
查询	since	否	string	按 updatedAt 的下界（含），RFC 3339。
查询	limit	否	integer	每页条数，1–200，默认 100。
查询	user	否	string	按调用时 user 标签精确筛选。
查询	apiKeyId	否	string	按调用所用 Account API Key 标识精确筛选（例如门户专用 Key）；只在主账号自己的调用范围内生效。
查询	apiKeyScope	否	string	portal_instance：只返回本门户实例历次门户专用 Key（用途 org_portal）产生的调用，含主账号变更前计在前任主账号名下的调用；仅门户专用 Key 可用，否则 403 portal_key_required。游标绑定门户实例，Key 轮换后仍有效。省略时为调用方 Key 所属工作空间的全部调用。

响应

HTTP 状态	说明	媒体类型与结构
200	请求成功	application/json：BillingCallList
400	请求参数不合法	application/json：PublicError

HTTP 状态	说明	媒体类型与结构
401	未认证或凭据失效	application/json：PublicError
402	额度不足	application/json：PublicError
403	没有访问权限	application/json：PublicError
404	资源不存在或不可见	application/json：PublicError
409	请求或状态冲突	application/json：PublicError
413	请求体超过大小限制	application/json：PublicError
415	不支持的媒体类型	application/json：PublicError
422	输入无法处理	application/json：PublicError
429	请求过于频繁	application/json：PublicError
500	服务内部错误	application/json：PublicError
502	上游响应异常	application/json：PublicError
503	服务暂不可用	application/json：PublicError
504	上游响应超时	application/json：PublicError

/docs/reference/models/createChatCompletion

生成对话回复

POST /v1/chat/completions

身份与入口

使用公开模型 API 基础地址和 Account API Key；不要混用应用 RuntimeSession。具体模型的条件字段由实时服务合同声明。

- 基础地址：/

认证：[{"bearerAuth":[]}]；凭据说明见类型附录。

请求参数

位置	参数	必填	类型	说明
请求头	X-Client-Request-Id	否	string	可选客户端关联值，只能包含可见 ASCII，最多 512 字节；不能替代服务端 requestId。

请求正文

请求正文必填。

application/json

类型：`ChatRequest`。完整字段见本文末尾的类型定义附录。

字段	类型	必填	说明
chat_template_kwargs	object	否	Qwen3.8 兼容的模板参数；建议使用顶层字段。同名字段不能与顶层冲突；json_schema 要求两处均关闭思考。
chat_template_kwargs.enable_thinking	boolean	否	仅支持思考的 Qwen3.8 服务：true 开启；与 reasoning_effort 同时传入时必须一致。默认关闭。；默认值：false
chat_template_kwargs.preserve_thinking	boolean	否	Qwen3.8 模板是否保留已有 assistant 思考历史；不是缓存或服务端会话开关。；默认值：true
chat_template_kwargs.reasoning_effort	string	否	Qwen3.8：none 关闭，minimal/low 映射 low，medium 映射 medium，high/xhigh/max 映射 xhigh；非 none 会开启思考。仅 enable_thinking=true 且省略此字段时采用 xhigh。不单独扩大输出预算。；可选值：["none","minimal","low","medium","high","xhigh","max"]

字段	类型	必填	说明
enable_thinking	boolean	否	仅支持思考的 Qwen3.8 服务：true 开启；与 reasoning_effort 同时传入时必须一致。默认关闭。；默认值：false
max_tokens	integer	否	本次输出预算，受模型输入输出合计上限约束；默认值由所选服务的冻结合同声明，包含思考与最终回答。思考请求应显式设置，预算耗尽返回 finish_reason=length，可能没有最终答案。；最小值：1
messages	array<object>	是	按对话顺序排列的消息集合。
messages[].content	string / array<object>	是	消息内容；支持字符串或模型允许的多模态片段。
messages[].reasoning_content	string / null	否	保留原始 assistant 消息中已有的思考历史，不拼入 user 正文。
messages[].role	string	是	消息或素材的角色，枚举含义依所属对象确定。；可选值：["system","user","assistant","tool"]
model	string	是	文本或视觉对话服务的公开模型别名。
preserve_thinking	boolean	否	Qwen3.8 模板是否保留已有 assistant 思考历史；不是缓存或服务端会话开关。；默认值：true
reasoning_effort	string	否	Qwen3.8：none 关闭，minimal/low 映射 low，medium 映射 medium，high/xhigh/max 映射 xhigh；非 none 会开启思考。仅 enable_thinking=true 且省略此字段时采用 xhigh。不单独扩大输出预算。；可选值：["none","minimal","low","medium","high","xhigh","max"]
stream	boolean	否	仅在服务支持流式输出时返回 SSE。采样参数须由模型合同允许。；默认值：false
stream_options	object	否	流式输出选项，仅在 stream=true 时使用。

字段	类型	必填	说明
stream_options.include_usage	boolean	否	流式结束前返回实际用量；该事件可能没有 choices。；默认值：false

响应

HTTP 状态	说明	媒体类型与结构
200	请求成功	application/json：ChatResponse；text/event-stream：string
400	请求参数不合法	application/json：PublicError
401	未认证或凭据失效	application/json：PublicError
402	额度不足	application/json：PublicError
403	没有访问权限	application/json：PublicError
404	资源不存在或不可见	application/json：PublicError
409	请求或状态冲突	application/json：PublicError
413	请求体超过大小限制	application/json：PublicError
415	不支持的媒体类型	application/json：PublicError
422	输入无法处理	application/json：PublicError
429	请求过于频繁	application/json：PublicError
500	服务内部错误	application/json：PublicError
502	上游响应异常	application/json：PublicError
503	服务暂不可用	application/json：PublicError
504	上游响应超时	application/json：PublicError

将文档解析为结构化文件

MinerU 异步文档解析。上传 PDF、DOCX/PPTX/XLSX 或 JPEG/PNG/WebP 后传入素材 ID；PDF 全文及 Office 必填 pages，page_range 仅接受 PDF 从 1 开始的闭区间。成功后遍历 outputs 下载全部文档产物，使用 manifest.artifacts 的角色与摘要区分两份 JSON；不开放取消，不使用媒体水印参数。产物默认以 personal_30d 保存到当前用户资源空间，自入库起 30 天到期，可在门户改为永久保存；顶层 storageMode=temporary_1h 仅后台已对用户或组织授权时允许，自入库起 1 小时到期。输入资源保留原期限，平台存储字段不传入解析模型。

`POST /v1/documents/parse`

身份与入口

使用公开模型 API 基础地址和 Account API Key；不要混用应用 RuntimeSession。具体模型的条件字段由实时服务合同声明。

- 基础地址：`/`

认证：[{"bearerAuth":[]}]；凭据说明见类型附录。

请求参数

位置	参数	必填	类型	说明
请求头	X-Client-Request-Id	否	string	可选客户端关联值，只能包含可见 ASCII，最多 512 字节；不能替代服务端 requestId。
请求头	Idempotency-Key	否	string	建议持久保存稳定业务幂等键；结果未知时使用原身份、原正文和原键恢复，已有 jobId 时查询原任务。

请求正文

请求正文必填。

application/json

类型：`DocumentParseRequest`。完整字段见本文末尾的类型定义附录。

字段	类型	必填	说明
input_document_asset_id	string	否	当前调用身份和模型范围内已上传的文档素材。
input_document_url	string	否	可选：与 input_document_asset_id 并列的外部来源，二选一。只接受公网 HTTPS（443 端口）；平台在接纳前经受控下载器摄入为当前调用身份和模型的一小时临时素材，再按资源 ID 原链路处理，等同先调用 /v1/assets/input。同槽位同时传 ID 与 URL 返回 400 ambiguous_input_source；同请求重复 URL 只下载一次；来源被拒为 400 invalid_source_url，下载失败 502 source_fetch_failed，超时 504 source_fetch_timeout；部署关闭本能力时返回 400 input_url_not_enabled。
language	string	否	语言代码；合法值由所选服务声明。；可选值：["ch","en"]
model	string	是	服务发现返回的公开模型别名。
original_name	string	否	原始文件名，不包含目录路径。
page_range	object	否	待处理的闭区间页码范围。
page_range.end_page	integer	是	结束页，闭区间；页码基准以所选解析服务为准。；最小值：1；最大值：500
page_range.start_page	integer	是	仅 PDF 接受页范围，从 1 开始的闭区间起点；Office 和图片不接受 page_range。；最小值：1；最大值：500
pages	integer	否	预占页数上限，范围 1–500；PDF 全文及 DOCX/PPTX/XLSX 必填，图片按一页。最终按实际 processedPages 结算。；最小值：1；最大值：500
preset	string	否	所选服务支持的解析预设。；可选值：["hybrid-medium","hybrid-high"]

字段	类型	必填	说明
storageMode	string	否	平台存储策略，使用顶层字段：默认 personal_30d，将本次主动上传或生成的资源保存到当前用户资源空间，自资源入库起 30 天到期，可在门户“我的资源”改为永久保存，仍占用配额。temporary_1h 自入库起 1 小时到期，仅后台已对所属用户或组织开启临时存储权限时可选；未授权时拒绝，不静默降级。生成请求不改变已引用素材的原到期时间。公网 URL 受控导入的临时输入遵循独立的 1 小时策略，不因本字段进入个人资源空间。；可选值：["personal_30d","temporary_1h"]；默认值："personal_30d"

响应

HTTP 状态	说明	媒体类型与结构
200	请求成功	application/json：DocumentParseJob
202	异步任务已受理或仍在执行；使用原 API Key 请求绝对 jobUrl（也见 Location），遵守 Retry-After 的查询间隔。	application/json：DocumentParseJob
202 响应头	Location	带鉴权的绝对轮询地址，与正文 jobUrl 相同。
202 响应头	Retry-After	建议距离下次轮询的等待秒数。
400	请求参数不合法	application/json：PublicError
401	未认证或凭据失效	application/json：PublicError
402	额度不足	application/json：PublicError
403	没有访问权限	application/json：PublicError
404	资源不存在或不可见	application/json：PublicError
409	请求或状态冲突	application/json：PublicError
413	请求体超过大小限制	application/json：PublicError
415	不支持的媒体类型	application/json：PublicError

HTTP 状态	说明	媒体类型与结构
422	输入无法处理	application/json：PublicError
429	请求过于频繁	application/json：PublicError
500	服务内部错误	application/json：PublicError
502	上游响应异常	application/json：PublicError
503	服务暂不可用	application/json：PublicError
504	上游响应超时	application/json：PublicError

200 · application/json 示例


```
{
  "id": "job_document_example",
  "jobId": "job_document_example",
  "jobUrl": "https://api.example.test/v1/jobs/job_document_example?model=mineru-document-parse",
  "model": "mineru-document-parse",
  "object": "document.parse.job",
  "requestId": "req_example",
  "serviceKind": "document_parse",
  "stage": "queued",
  "status": "queued"
}
```


202 · application/json 示例


```
{
  "id": "job_document_example",
  "jobId": "job_document_example",
  "jobUrl": "https://api.example.test/v1/jobs/job_document_example?model=mineru-document-parse",
  "model": "mineru-document-parse",
  "object": "document.parse.job",
  "requestId": "req_example",
  "serviceKind": "document_parse",
  "stage": "queued",
  "status": "queued"
}
```


/docs/reference/models/createEmbedding

生成文本向量

POST /v1/embeddings

身份与入口

使用公开模型 API 基础地址和 Account API Key；不要混用应用 RuntimeSession。具体模型的条件字段由实时服务合同声明。

- 基础地址：`/`

认证：[{"bearerAuth":[]}]；凭据说明见类型附录。

请求参数

位置	参数	必填	类型	说明
请求头	X-Client-Request-Id	否	string	可选客户端关联值，只能包含可见 ASCII，最多 512 字节；不能替代服务端 requestId。

请求正文

请求正文必填。

application/json

类型：`EmbeddingRequest`。完整字段见本文末尾的类型定义附录。

字段	类型	必填	说明
input	string / array<string>	是	去除首尾空白后不可为空；最多 64 个非空文本，每项最多 8192 个 Unicode 字符。
model	string	是	向量模型的公开别名。

响应

HTTP 状态	说明	媒体类型与结构
200	请求成功	application/json：EmbeddingResponse
400	请求参数不合法	application/json：PublicError
401	未认证或凭据失效	application/json：PublicError
402	额度不足	application/json：PublicError

HTTP 状态	说明	媒体类型与结构
403	没有访问权限	application/json：PublicError
404	资源不存在或不可见	application/json：PublicError
409	请求或状态冲突	application/json：PublicError
413	请求体超过大小限制	application/json：PublicError
415	不支持的媒体类型	application/json：PublicError
422	输入无法处理	application/json：PublicError
429	请求过于频繁	application/json：PublicError
500	服务内部错误	application/json：PublicError
502	上游响应异常	application/json：PublicError
503	服务暂不可用	application/json：PublicError
504	上游响应超时	application/json：PublicError

/docs/reference/models/createImageGeneration

生成图片

生成产物默认以 personal_30d 入库当前用户资源空间，占用空间配额，自入库起 30 天到期；可在门户改为永久保存。顶层 storageMode=temporary_1h 仅后台已对用户或组织授权时允许，自入库起 1 小时到期。引用既有资源不会改变其期限；公网 URL 临时输入继续按独立 1 小时策略处理。普通 API Key 不接受应用来源或隐藏产物字段。

`POST /v1/images/generations`

身份与入口

使用公开模型 API 基础地址和 Account API Key；不要混用应用 RuntimeSession。具体模型的条件字段由实时服务合同声明。

- 基础地址：/

认证：[{"bearerAuth":[]}]；凭据说明见类型附录。

请求参数

位置	参数	必填	类型	说明
请求头	X-Client-Request-Id	否	string	可选客户端关联值，只能包含可见 ASCII，最多 512 字节；不能替代服务端 requestId。

请求正文

请求正文必填。

application/json

类型：`ImageGenerationRequest`。完整字段见本文末尾的类型定义附录。

字段	类型	必填	说明
aspectRatio	string	否	受控输出长宽比；H3 由服务能力合同返回，例如 21:9、16:9、4:3、1:1、3:4、9:16。
generationMode	string	否	模型支持的生成模式，例如 t2i（文生图）、i2i（图生图）。
imageCount	integer	否	请求生成图片数量；以模型允许范围为准。；最小值：1
input_image_asset_id	string	否	受控单张图片素材，仅用于支持图片输入的模型。
input_image_asset_ids	array<string>	否	按顺序排列的输入图片 ID；仅支持多图的模型可用，不与单图字段混传。

字段	类型	必填	说明
input_image_url	string	否	可选：与 input_image_asset_id 并列的外部来源，二选一。只接受公网 HTTPS（443 端口），另接受 图片数据；平台在接纳前经受控下载器摄入为当前调用身份和模型的一小时临时素材，再按资源 ID 原链路处理，等同先调用 /v1/assets/input。同槽位同时传 ID 与 URL 返回 400 ambiguous_input_source；同请求重复 URL 只下载一次；来源被拒为 400 invalid_source_url，下载失败 502 source_fetch_failed，超时 504 source_fetch_timeout；部署关闭本能力时返回 400 input_url_not_enabled。
input_image_urls	array<string>	否	有序 URL 数组，按位置改写为 input_image_asset_ids。可选：与 input_image_asset_ids 并列的外部来源，二选一。只接受公网 HTTPS（443 端口），另接受 图片数据；平台在接纳前经受控下载器摄入为当前调用身份和模型的一小时临时素材，再按资源 ID 原链路处理，等同先调用 /v1/assets/input。同槽位同时传 ID 与 URL 返回 400 ambiguous_input_source；同请求重复 URL 只下载一次；来源被拒为 400 invalid_source_url，下载失败 502 source_fetch_failed，超时 504 source_fetch_timeout；部署关闭本能力时返回 400 input_url_not_enabled。
model	string	是	图片模型的公开别名。
orientation	string	否	画幅方向，合法值按模型规格选择。

字段	类型	必填	说明
outputSizeCode	string	否	模型公布的输出尺寸代码；不要自行拼接尺寸。
prompt	string	是	图片生成提示词；字符上限由所选服务声明。
qualityTier	string	否	模型公布的质量档位。
reference_image_url	string	否	可选：与 reference_image_asset_id 并列的外部来源，二选一。只接受公网 HTTPS（443 端口），另接受 图片数据；平台在接纳前经受控下载器摄入为当前调用身份和模型的一小时临时素材，再按资源 ID 原链路处理，等同先调用 /v1/assets/input。同槽位同时传 ID 与 URL 返回 400 ambiguous_input_source；同请求重复 URL 只下载一次；来源被拒为 400 invalid_source_url，下载失败 502 source_fetch_failed，超时 504 source_fetch_timeout；部署关闭本能力时返回 400 input_url_not_enabled。

字段	类型	必填	说明
reference_image_urls	array<string>	否	有序 URL 数组，按位置改写为 reference_image_asset_ids。可选：与 reference_image_asset_ids 并列的外部来源，二选一。只接受公网 HTTPS（443 端口），另接受 图片数据；平台在接纳前经受控下载器摄入为当前调用身份和模型的一小时临时素材，再按资源 ID 原链路处理，等同先调用 /v1/assets/input。同槽位同时传 ID 与 URL 返回 400 ambiguous_input_source；同请求重复 URL 只下载一次；来源被拒为 400 invalid_source_url，下载失败 502 source_fetch_failed，超时 504 source_fetch_timeout；部署关闭本能力时返回 400 input_url_not_enabled。
resolutionTier	string	否	所选模型支持的清晰度档位。
seed	integer	否	可选随机种子；幂等重试时保持原值。
size	string	否	图片尺寸，格式为宽x高，例如 1024x1024；仅可使用所选服务 mediaSpecContract 声明的尺寸。

字段	类型	必填	说明
storageMode	string	否	平台存储策略，使用顶层字段：默认 personal_30d，将本次主动上传或生成的资源保存到当前用户资源空间，自资源入库起 30 天到期，可在门户“我的资源”改为永久保存，仍占用配额。temporary_1h 自入库起 1 小时到期，仅后台已对所属用户或组织开启临时存储权限时可选；未授权时拒绝，不静默降级。生成请求不改变已引用素材的原到期时间。公网 URL 受控导入的临时输入遵循独立的 1 小时策略，不因本字段进入个人资源空间。；可选值：["personal_30d","temporary_1h"]；默认值："personal_30d"

响应

HTTP 状态	说明	媒体类型与结构
200	返回重放的异步任务，或同步执行服务的结果。H3 使用 MediaJob 分支；请核对所选服务的执行方式。	application/json：MediaJob / SynchronousMediaResponse
202	异步任务已受理或仍在执行；使用原 API Key 请求绝对 jobUrl（也见 Location），遵守 Retry-After 的查询间隔。	application/json：MediaJob
202 响应头	Location	带鉴权的绝对轮询地址，与正文 jobUrl 相同。
202 响应头	Retry-After	建议距离下次轮询的等待秒数。
400	请求参数不合法	application/json：PublicError
401	未认证或凭据失效	application/json：PublicError
402	额度不足	application/json：PublicError
403	没有访问权限	application/json：PublicError
404	资源不存在或不可见	application/json：PublicError
409	请求或状态冲突	application/json：PublicError
413	请求体超过大小限制	application/json：PublicError

HTTP 状态	说明	媒体类型与结构
415	不支持的媒体类型	application/json：PublicError
422	输入无法处理	application/json：PublicError
429	请求过于频繁	application/json：PublicError
500	服务内部错误	application/json：PublicError
502	上游响应异常	application/json：PublicError
503	服务暂不可用	application/json：PublicError
504	上游响应超时	application/json：PublicError

202 · application/json 示例


```
{
  "id": "job_example",
  "jobId": "job_example",
  "jobUrl": "https://api.example.test/v1/jobs/job_example?model=minimax-h3",
  "model": "minimax-h3",
  "requestId": "req_example",
  "serviceKind": "video_generation",
  "stage": "queued",
  "status": "queued"
}
```


/docs/reference/models/getJob

查询异步任务

GET /v1/jobs/{id}

身份与入口

使用公开模型 API 基础地址和 Account API Key；不要混用应用 RuntimeSession。具体模型的条件字段由实时服务合同声明。

- 基础地址：/

认证：[{"bearerAuth":[]}]；凭据说明见类型附录。

请求参数

位置	参数	必填	类型	说明
路径	id	是	string	
查询	model	是	string	限定任务或素材所属模型范围的公开别名；须在当前 API Key 授权范围内。
请求头	X-Client-Request-Id	否	string	可选客户端关联值，只能包含可见 ASCII，最多 512 字节；不能替代服务端 requestId。

响应

HTTP 状态	说明	媒体类型与结构
200	任务快照。202 仅表示受理或处理中；读取 status 判断终态，重放可能返回已结束任务。H3 不提供用户取消 API。	application/json：MediaJob
202	任务快照。202 仅表示受理或处理中；读取 status 判断终态，重放可能返回已结束任务。H3 不提供用户取消 API。	application/json：MediaJob
202 响应头	Location	带鉴权的绝对轮询地址，与正文 jobUrl 相同。
202 响应头	Retry-After	建议距离下次轮询的等待秒数。
400	请求参数不合法	application/json：PublicError
401	未认证或凭据失效	application/json：PublicError
402	额度不足	application/json：PublicError
403	没有访问权限	application/json：PublicError
404	资源不存在或不可见	application/json：PublicError
409	请求或状态冲突	application/json：PublicError
413	请求体超过大小限制	application/json：PublicError
415	不支持的媒体类型	application/json：PublicError
422	输入无法处理	application/json：PublicError

HTTP 状态	说明	媒体类型与结构
429	请求过于频繁	application/json：PublicError
500	服务内部错误	application/json：PublicError
502	上游响应异常	application/json：PublicError
503	服务暂不可用	application/json：PublicError
504	上游响应超时	application/json：PublicError

200 · application/json 示例


```
{
  "id": "job_example",
  "jobId": "job_example",
  "jobUrl": "https://api.example.test/v1/jobs/job_example?model=minimax-h3",
  "model": "minimax-h3",
  "requestId": "req_example",
  "serviceKind": "video_generation",
  "stage": "queued",
  "status": "queued"
}
```


202 · application/json 示例


```
{
  "id": "job_example",
  "jobId": "job_example",
  "jobUrl": "https://api.example.test/v1/jobs/job_example?model=minimax-h3",
  "model": "minimax-h3",
  "requestId": "req_example",
  "serviceKind": "video_generation",
  "stage": "queued",
  "status": "queued"
}
```

生成视频

创建可能计费的异步视频任务。公开 Account API Key H3 四模式缺省自动优化，显式 optimizePrompt=false 跳过且不收优化费；已手动优化时须显式关闭。后台最多两次常规优化、每次最多一次语义压缩，最多四次模型调用按实际 token 单独计费，全部计入 Job 30 分钟总预算。优化失败或安全拒绝不会回退原文生成，失败查询返回 prompt_optimization_failed；未知结果只恢复同一 attempt。保存原正文、幂等键和 jobId，受理未知只用原身份/原正文/原键恢复，已有 jobId 则查询原任务。其他视频模型无默认优化。 生成产物默认以 personal_30d 入库当前用户资源空间，占用空间配额，自入库起 30 天到期；可在门户改为永久保存。顶层 storageMode=temporary_1h 仅后台已对用户或组织授权时允许，自入库起 1 小时到期。引用既有资源不会改变其期限；公网 URL 临时输入继续按独立 1 小时策略处理。普通 API Key 不接受应用来源或隐藏产物字段。

`POST /v1/media/generations`

身份与入口

使用公开模型 API 基础地址和 Account API Key；不要混用应用 RuntimeSession。具体模型的条件字段由实时服务合同声明。

- 基础地址：`/`

认证：[{"bearerAuth":[]}]；凭据说明见类型附录。

请求参数

位置	参数	必填	类型	说明
请求头	X-Client-Request-Id	否	string	可选客户端关联值，只能包含可见 ASCII，最多 512 字节；不能替代服务端 requestId。
请求头	Idempotency-Key	否	string	强烈建议提供稳定业务幂等键，以便在结果未知时安全重放。同键用于不同规范请求会发生冲突。

请求正文

请求正文必填。

application/json

类型：`MediaGenerationRequest`。完整字段见本文末尾的类型定义附录。

公开视频接口使用顶层字段，不使用正式应用的 params/mediaSpec 包装。H3 必须提供 generationMode、resolutionTier、orientation、seconds，支持 544p/768p、六种受控画布和 5 ~ 15 秒，提示词最多 7000 字符。t2v 不传素材；i2v 传首图；首尾帧传两张图；全能参考传分类素材。其他模型的条件必填和合法组合由 mediaSpecContract 声明；不要向 H3 传 fps、frames、width、height、steps 或运行时设置。

字段	类型	必填	说明
aspectRatio	string	否	H3 受控画面比例，例如 21:9、16:9、4:3、1:1、3:4、9:16；需与 outputSizeCode 配套。
characterMode	string	否	Wan2.2 Animate 专用：动作迁移或角色替换；需要人物图和驱动视频，不传 prompt。其他视频模型不使用本字段。；可选值：["motion_transfer","character_replace"]
enableRife	boolean	否	仅部分普通 Wan 视频服务支持的插帧选项；H3、Wan Animate 和 X-Dub 不接受。须先确认所选服务能力。
end_image_asset_id	string	否	已上传的尾帧图片；仅用于 H3 首尾帧等明确支持尾帧的模式。
end_image_url	string	否	可选：与 end_image_asset_id 并列的外部来源，二选一。只接受公网 HTTPS（443 端口），另接受 图片数据；平台在接纳前经受控下载器摄入为当前调用身份和模型的一小时临时素材，再按资源 ID 原链路处理，等同先调用 /v1/assets/input。同槽位同时传 ID 与 URL 返回 400 ambiguous_input_source；同请求重复 URL 只下载一次；来源被拒为 400 invalid_source_url，下载失败 502 source_fetch_failed，超时 504 source_fetch_timeout；部署关闭本能力时返回 400 input_url_not_enabled。

字段	类型	必填	说明
generationMode	string	否	H3 必填，可选 t2v、i2v、first_last_frame_video、universal_reference_video；其他模型使用各自公布的模式。
input_audio_asset_id	string	否	明确支持音频输入的模型所需素材；不是 H3 全能参考的独立顶层字段。
input_audio_url	string	否	可选：与 input_audio_asset_id 并列的外部来源，二选一。只接受公网 HTTPS（443 端口）；平台在接纳前经受控下载器摄入为当前调用身份和模型的一小时临时素材，再按资源 ID 原链路处理，等同先调用 /v1/assets/input。同槽位同时传 ID 与 URL 返回 400 ambiguous_input_source；同请求重复 URL 只下载一次；来源被拒为 400 invalid_source_url，下载失败 502 source_fetch_failed，超时 504 source_fetch_timeout；部署关闭本能力时返回 400 input_url_not_enabled。
input_image_asset_id	string	否	已上传的输入图片；H3 用于图生和首尾帧，Wan Animate 用于人物图；其他模型按服务合同使用。
input_image_asset_ids	array<string>	否	有序输入图片 ID；Bernini r2v 使用 1 ~ 5 张，rv2v 使用恰好 1 张。不是 H3 全能参考字段；仅按当前模型能力使用。

字段	类型	必填	说明
input_image_url	string	否	可选：与 input_image_asset_id 并列的外部来源，二选一。只接受公网 HTTPS（443 端口），另接受 图片数据；平台在接纳前经受控下载器摄入为当前调用身份和模型的一小时临时素材，再按资源 ID 原链路处理，等同先调用 /v1/assets/input。同槽位同时传 ID 与 URL 返回 400 ambiguous_input_source；同请求重复 URL 只下载一次；来源被拒为 400 invalid_source_url，下载失败 502 source_fetch_failed，超时 504 source_fetch_timeout；部署关闭本能力时返回 400 input_url_not_enabled。
input_image_urls	array<string>	否	有序 URL 数组，按位置改写为 input_image_asset_ids。可选：与 input_image_asset_ids 并列的外部来源，二选一。只接受公网 HTTPS（443 端口），另接受 图片数据；平台在接纳前经受控下载器摄入为当前调用身份和模型的一小时临时素材，再按资源 ID 原链路处理，等同先调用 /v1/assets/input。同槽位同时传 ID 与 URL 返回 400 ambiguous_input_source；同请求重复 URL 只下载一次；来源被拒为 400 invalid_source_url，下载失败 502 source_fetch_failed，超时 504 source_fetch_timeout；部署关闭本能力时返回 400 input_url_not_enabled。
input_video_asset_id	string	否	明确支持视频输入的模型所需素材；不是 H3 全能参考的独立顶层字段。

字段	类型	必填	说明
input_video_url	string	否	可选：与 input_video_asset_id 并列的外部来源，二选一。只接受公网 HTTPS（443 端口）；平台在接纳前经受控下载器摄入为当前调用身份和模型的一小时临时素材，再按资源 ID 原链路处理，等同先调用 /v1/assets/input。同槽位同时传 ID 与 URL 返回 400 ambiguous_input_source；同请求重复 URL 只下载一次；来源被拒为 400 invalid_source_url，下载失败 502 source_fetch_failed，超时 504 source_fetch_timeout；部署关闭本能力时返回 400 input_url_not_enabled。
model	string	是	GET /v1/services 返回的公开模型别名；不能使用部署名或工作负载名。
optimizePrompt	boolean	否	仅公开 Account API Key H3 四模式生成缺省开启。部署未激活时，缺省或 true 返回 503 prompt_optimization_capability_unavailable；查询 automaticPromptOptimization.available 获取编排可用性，它不代表模型实时健康。显式 false 完全跳过优化且不收优化费；已手动优化的 prompt 应传 false。其他模型不会自动优化，true 返回参数错误。后台准备最多两次常规优化，每次完整超长候选最多压缩一次，最多四次模型调用按实际 token 单独计费。两次失败则 Job 返回 prompt_optimization_failed，不回退原 prompt；未知结果恢复原 attempt，不新发模型调用。
orientation	string	否	画幅方向，合法值按模型规格选择。；可选值: ["landscape","portrait"]
outputSizeCode	string	否	H3 受控画布代码；服务能力合同返回的六种 544p 或六种 768p 代码之一。

字段	类型	必填	说明
prompt	string	否	生成提示词；H3 去除首尾空白后为 1 ~ 7000 个 Unicode 字符。Wan Animate、X-Dub 等无提示词模式不传。
referenceInputs	array<H3GenerationReference>	否	H3 全能参考：无视频最多 9 图，有 1 个视频时最多 8 图；最多 3 段音频，总计不超过 12 项且至少一项视觉素材。同类型顺序决定编号；素材 ID 不可重复，且须属于当前调用身份和模型。；最多项数：12
reference_image_url	string	否	可选：与 reference_image_asset_id 并列的外部来源，二选一。只接受公网 HTTPS（443 端口），另接受 图片数据；平台在接纳前经受控下载器摄入为当前调用身份和模型的一小时临时素材，再按资源 ID 原链路处理，等同先调用 /v1/assets/input。同槽位同时传 ID 与 URL 返回 400 ambiguous_input_source；同请求重复 URL 只下载一次；来源被拒为 400 invalid_source_url，下载失败 502 source_fetch_failed，超时 504 source_fetch_timeout；部署关闭本能力时返回 400 input_url_not_enabled。

字段	类型	必填	说明
reference_image_urls	array<string>	否	有序 URL 数组，按位置改写为 reference_image_asset_ids。可选：与 reference_image_asset_ids 并列的外部来源，二选一。只接受公网 HTTPS（443 端口），另接受 图片数据；平台在接纳前经受控下载器摄入为当前调用身份和模型的一小时临时素材，再按资源 ID 原链路处理，等同先调用 /v1/assets/input。同槽位同时传 ID 与 URL 返回 400 ambiguous_input_source；同请求重复 URL 只下载一次；来源被拒为 400 invalid_source_url，下载失败 502 source_fetch_failed，超时 504 source_fetch_timeout；部署关闭本能力时返回 400 input_url_not_enabled。
resolutionTier	string	否	H3 支持 544p、768p；其他服务使用各自公布的清晰度。
seconds	integer	否	请求生成时长，单位秒。H3 为 5 ~ 15 的整数；实际产物时长和帧数另由结果确定。
seed	integer	否	可选随机种子，用于可重复输入；同一幂等请求重试时不得改变。；最小值：0；最大值：2147483647

字段	类型	必填	说明
source_video_url	string	否	可选：与 source_video_asset_id 并列的外部来源，二选一。只接受公网 HTTPS（443 端口）；平台在接纳前经受控下载器摄入为当前调用身份和模型的一小时临时素材，再按资源 ID 原链路处理，等同先调用 /v1/assets/input。同槽位同时传 ID 与 URL 返回 400 ambiguous_input_source；同请求重复 URL 只下载一次；来源被拒为 400 invalid_source_url，下载失败 502 source_fetch_failed，超时 504 source_fetch_timeout；部署关闭本能力时返回 400 input_url_not_enabled。
speedMultiplier	number	否	Wan2.2 Animate 处理速度。720p 支持 1 或 1.5；480p 使用 1。1.5 倍按对应价格系数计费，不承诺上传和排队等总耗时同比缩短。其他模型仅在能力合同声明时使用。；可选值：[1,1.5]
storageMode	string	否	平台存储策略，使用顶层字段：默认 personal_30d，将本次主动上传或生成的资源保存到当前用户资源空间，自资源入库起 30 天到期，可在门户“我的资源”改为永久保存，仍占用配额。temporary_1h 自入库起 1 小时到期，仅后台已对所属用户或组织开启临时存储权限时可选；未授权时拒绝，不静默降级。生成请求不改变已引用素材的原到期时间。公网 URL 受控导入的临时输入遵循独立的 1 小时策略，不因本字段进入个人资源空间。；可选值：["personal_30d","temporary_1h"]；默认值："personal_30d"

```
{
  "generationMode": "t2v",
  "model": "minimax-h3",
  "orientation": "landscape",
  "prompt": "雨夜站台，镜头缓慢推进，人物自然递伞后共同离开。",
  "resolutionTier": "768p",
  "seconds": 5,
  "seed": 1234,
  "storageMode": "personal_30d"
}
```

```
{
  "generationMode": "i2v",
  "input_image_asset_id": "input_first",
  "model": "minimax-h3",
  "orientation": "landscape",
  "prompt": "雨夜站台，镜头缓慢推进，人物自然递伞后共同离开。",
  "resolutionTier": "768p",
  "seconds": 5,
  "seed": 1234,
  "storageMode": "personal_30d"
}
```

```
{
  "end_image_asset_id": "input_last",
  "generationMode": "first_last_frame_video",
  "input_image_asset_id": "input_first",
  "model": "minimax-h3",
  "orientation": "landscape",
  "prompt": "雨夜站台，镜头缓慢推进，人物自然递伞后共同离开。",
  "resolutionTier": "768p",
  "seconds": 5,
  "seed": 1234,
  "storageMode": "personal_30d"
}
```

```
{
  "generationMode": "universal_reference_video",
  "model": "minimax-h3",
  "orientation": "landscape",
  "prompt": "雨夜站台，镜头缓慢推进，人物自然递伞后共同离开。",
  "referenceInputs": [
    {
      "assetId": "input_storyboard",
      "mediaType": "image",
      "role": "reference_image"
    }
  ],
  "resolutionTier": "768p",
  "seconds": 5,
  "seed": 1234,
  "storageMode": "personal_30d"
}
```

响应

HTTP 状态	说明	媒体类型与结构
200	返回重放的异步任务，或同步执行服务的结果。H3 使用 MediaJob 分支；请核对所选服务的执行方式。	application/json：MediaJob / SynchronousMediaResponse
202	任务快照。202 仅表示受理或处理中；读取 status 判断终态，重放可能返回已结束任务。H3 不提供用户取消 API。	application/json：MediaJob
202 响应头	Location	带鉴权的绝对轮询地址，与正文 jobUrl 相同。
202 响应头	Retry-After	建议距离下次轮询的等待秒数。
400	请求参数不合法	application/json：PublicError
401	未认证或凭据失效	application/json：PublicError
402	额度不足	application/json：PublicError
403	没有访问权限	application/json：PublicError
404	资源不存在或不可见	application/json：PublicError
409	请求或状态冲突	application/json：PublicError
413	请求体超过大小限制	application/json：PublicError
415	不支持的媒体类型	application/json：PublicError
422	输入无法处理	application/json：PublicError

HTTP 状态	说明	媒体类型与结构
429	请求过于频繁	application/json：PublicError
500	服务内部错误	application/json：PublicError
502	上游响应异常	application/json：PublicError
503	服务暂不可用	application/json：PublicError
504	上游响应超时	application/json：PublicError

202 · application/json 示例


```
{
  "id": "job_example",
  "jobId": "job_example",
  "jobUrl": "https://api.example.test/v1/jobs/job_example?model=minimax-h3",
  "model": "minimax-h3",
  "requestId": "req_example",
  "serviceKind": "video_generation",
  "stage": "queued",
  "status": "queued"
}
```


/docs/reference/models/createPromptOptimization

优化媒体提示词

POST /v1/media/prompt-optimizations

身份与入口

使用公开模型 API 基础地址和 Account API Key；不要混用应用 RuntimeSession。具体模型的条件字段由实时服务合同声明。

- 基础地址：/

认证：[{"bearerAuth":[]}]；凭据说明见类型附录。

请求参数

位置	参数	必填	类型	说明
请求头	X-Client-Request-Id	否	string	可选客户端关联值，只能包含可见 ASCII，最多 512 字节；不能替代服务端 requestId。
请求头	Idempotency-Key	是	string	必须与 clientRequestId 完全相同；网络结果未知时使用相同正文和相同键恢复。

请求正文

请求正文必填。

application/json

类型：`PromptOptimizationRequest`。完整字段见本文末尾的类型定义附录。

字段	类型	必填	说明
audioReferencePolicy	string	否	H3 存在音频参考时必填。优化器只理解图片/视频；音频仅按明确文字意图保留用途，不分析其内容。；可选值：["preserve_without_understanding"]
clientRequestId	string	是	客户端业务请求标识；优化操作中须与幂等键一致。；最短长度：8；最长长度：128；匹配规则：^[A-Za-z0-9._:-]+$
dialogueLanguage	string	否	可选台词语种。auto（默认）表示由提示词语言自动解析；其余值为官方稳定支持的台词语言。分角色语种请在提示词中写明。仅 H3 typed-media/v1 支持。；可选值：["auto","arabic","chinese","english","french","german","italian","japanese","korean","portuguese","russian","spanish"]

字段	类型	必填	说明
generationMode	string	是	允许的模式来自所选服务的 promptOptimization.profiles。
mediaSpec	PromptOptimizationMediaSpec	是	类型定义：PromptOptimizationMediaSpec
model	string	是	GET /v1/services 返回的公开模型别名。
originalPrompt	string	是	Qwen Image 2.1 支持最多 10000 个 Unicode 字符；H3 typed-media/v1 仍最多 7000，其他优化配置按所选服务公布的字符上限。；最短长度：1；最长长度：10000
referenceInputs	array<object>	是	使用当前身份上传的素材 ID。H3 角色为 first_frame/last_frame，或 reference_image（0～9）、reference_video（0～1）、reference_audio（0～3）；全能参考至少一项视觉素材，总计不超过 12。同类型顺序决定编号。Qwen Image 2.1 文生图为空数组，编辑为有序 reference_image_1..reference_image_10（1～10 张）；其他配置使用各自角色合同。；最多项数：12
referenceInputs[].assetId	string	是	素材唯一标识，不是任务 ID。
referenceInputs[].role	string	是	消息或素材的角色，枚举含义依所属对象确定。
style	string	否	H3 typed-media/v1 的可选创作风格偏好。；最短长度：1；最长长度：200

响应

HTTP 状态	说明	媒体类型与结构
201	提示词优化完成并已保存	application/json：PromptOptimizationResponse
400	请求参数不合法	application/json：PublicError

HTTP 状态	说明	媒体类型与结构
401	未认证或凭据失效	application/json：PublicError
402	额度不足	application/json：PublicError
403	没有访问权限	application/json：PublicError
404	资源不存在或不可见	application/json：PublicError
409	请求或状态冲突	application/json：PublicError
413	请求体超过大小限制	application/json：PublicError
415	不支持的媒体类型	application/json：PublicError
422	输入无法处理	application/json：PublicError
429	请求过于频繁	application/json：PublicError
500	服务内部错误	application/json：PublicError
502	上游响应异常	application/json：PublicError
503	服务暂不可用	application/json：PublicError
504	上游响应超时	application/json：PublicError

/docs/reference/models/listModels

查询可用模型别名

GET /v1/models

身份与入口

使用公开模型 API 基础地址和 Account API Key；不要混用应用 RuntimeSession。具体模型的条件字段由实时服务合同声明。

- 基础地址：/

认证：[{"bearerAuth":[]}]；凭据说明见类型附录。

请求参数

位置	参数	必填	类型	说明
请求头	X-Client-Request-Id	否	string	可选客户端关联值，只能包含可见 ASCII，最多 512 字节；不能替代服务端 requestId。

响应

HTTP 状态	说明	媒体类型与结构
200	请求成功	application/json：PublicModelList
400	请求参数不合法	application/json：PublicError
401	未认证或凭据失效	application/json：PublicError
402	额度不足	application/json：PublicError
403	没有访问权限	application/json：PublicError
404	资源不存在或不可见	application/json：PublicError
409	请求或状态冲突	application/json：PublicError
413	请求体超过大小限制	application/json：PublicError
415	不支持的媒体类型	application/json：PublicError
422	输入无法处理	application/json：PublicError
429	请求过于频繁	application/json：PublicError
500	服务内部错误	application/json：PublicError
502	上游响应异常	application/json：PublicError
503	服务暂不可用	application/json：PublicError
504	上游响应超时	application/json：PublicError

识别文档文字

POST /v1/ocr

身份与入口

使用公开模型 API 基础地址和 Account API Key；不要混用应用 RuntimeSession。具体模型的条件字段由实时服务合同声明。

- 基础地址：/

认证：[{"bearerAuth":[]}]；凭据说明见类型附录。

请求参数

位置	参数	必填	类型	说明
请求头	X-Client-Request-Id	否	string	可选客户端关联值，只能包含可见 ASCII，最多 512 字节；不能替代服务端 requestId。

请求正文

请求正文必填。

application/json

类型：OCRRequest。完整字段见本文末尾的类型定义附录。

字段	类型	必填	说明
document_asset_id	string	否	仅用于明确支持该字段的 OCR 服务；不同于文档解析的素材字段。
image_asset_id	string	否	计量兼容的输入素材标识。

字段	类型	必填	说明
input_image_asset_id	string	否	图片 OCR 推荐使用 /assets/input 上传所得图片 ID；是否必填由具体模型决定。
input_image_url	string	否	可选：与 input_image_asset_id / image_asset_id / document_asset_id 并列的外部来源，二选一。只接受公网 HTTPS（443 端口），另接受 图片数据；平台在接纳前经受控下载器摄入为当前调用身份和模型的一小时临时素材，再按资源 ID 原链路处理，等同先调用 /v1/assets/input。同槽位同时传 ID 与 URL 返回 400 ambiguous_input_source；同请求重复 URL 只下载一次；来源被拒为 400 invalid_source_url，下载失败 502 source_fetch_failed，超时 504 source_fetch_timeout；部署关闭本能力时返回 400 input_url_not_enabled。
model	string	是	OCR 服务公开别名；实际图片格式及任务字段由模型能力决定。
pages	integer	否	计量页数；省略或非正数时默认为 1，不表示允许任意指定执行页数。
task	string	否	模型支持的识别任务，例如 ocr。

响应

HTTP 状态	说明	媒体类型与结构
200	请求成功	application/json：OCRResponse
400	请求参数不合法	application/json：PublicError
401	未认证或凭据失效	application/json：PublicError

HTTP 状态	说明	媒体类型与结构
402	额度不足	application/json：PublicError
403	没有访问权限	application/json：PublicError
404	资源不存在或不可见	application/json：PublicError
409	请求或状态冲突	application/json：PublicError
413	请求体超过大小限制	application/json：PublicError
415	不支持的媒体类型	application/json：PublicError
422	输入无法处理	application/json：PublicError
429	请求过于频繁	application/json：PublicError
500	服务内部错误	application/json：PublicError
502	上游响应异常	application/json：PublicError
503	服务暂不可用	application/json：PublicError
504	上游响应超时	application/json：PublicError

/docs/reference/models/getOpenAPI

读取公开 API 协议

GET /v1/openapi.json

身份与入口

使用公开模型 API 基础地址和 Account API Key；不要混用应用 RuntimeSession。具体模型的条件字段由实时服务合同声明。

- 基础地址：/

认证：[]；凭据说明见类型附录。

请求参数

位置	参数	必填	类型	说明

响应

HTTP 状态	说明	媒体类型与结构
200	根据已安装的公开路由生成的 OpenAPI 3.1 文档	application/json：OpenAPIDocument
400	请求参数不合法	application/json：PublicError
401	未认证或凭据失效	application/json：PublicError
402	额度不足	application/json：PublicError
403	没有访问权限	application/json：PublicError
404	资源不存在或不可见	application/json：PublicError
409	请求或状态冲突	application/json：PublicError
413	请求体超过大小限制	application/json：PublicError
415	不支持的媒体类型	application/json：PublicError
422	输入无法处理	application/json：PublicError
429	请求过于频繁	application/json：PublicError
500	服务内部错误	application/json：PublicError
502	上游响应异常	application/json：PublicError
503	服务暂不可用	application/json：PublicError
504	上游响应超时	application/json：PublicError

/docs/reference/models/getOrganizationCompute

查询组织专属算力

返回组织专属集群（集群共享策略为按组织绑定且仅绑定本组织）的卡型卡数、利用率、显存、服务副本、剩余容量、节点健康与观测时间；没有专属集群返回空 data，门户据此隐藏“专属算力”菜单。开发环境调试观测把选中的共用集群按同一投影返回并带 debug=true。身份规则同组织查询接口；观测源不可读返回 503 compute_unavailable（可重试），不以空列表冒充。

GET /v1/organization/compute

身份与入口

使用公开模型 API 基础地址和 Account API Key；不要混用应用 RuntimeSession。具体模型的条件字段由实时服务合同声明。

- 基础地址：/

认证：[{"bearerAuth":[]}]；凭据说明见类型附录。

请求参数

位置	参数	必填	类型	说明
请求头	X-Client-Request-Id	否	string	可选客户端关联值，只能包含可见 ASCII，最多 512 字节；不能替代服务端 requestId。

响应

HTTP 状态	说明	媒体类型与结构
200	请求成功	application/json：OrganizationCompute
400	请求参数不合法	application/json：PublicError
401	未认证或凭据失效	application/json：PublicError
402	额度不足	application/json：PublicError
403	没有访问权限	application/json：PublicError
404	资源不存在或不可见	application/json：PublicError
409	请求或状态冲突	application/json：PublicError
413	请求体超过大小限制	application/json：PublicError
415	不支持的媒体类型	application/json：PublicError
422	输入无法处理	application/json：PublicError
429	请求过于频繁	application/json：PublicError
500	服务内部错误	application/json：PublicError
502	上游响应异常	application/json：PublicError

HTTP 状态	说明	媒体类型与结构
503	服务暂不可用	application/json：PublicError
504	上游响应超时	application/json：PublicError

/docs/reference/models/getOrganizationContext

查询组织与当前 Key 身份

仅组织当前 active 主账号的 Account API Key 可调用：非组织成员返回 403 organization_membership_required，非主账号返回 403 organization_primary_required，组织停用返回 403 organization_access_suspended，组织上下文不可用返回 503 organization_context_unavailable。主账号身份每次请求实时校验。

`GET /v1/organization/context`

身份与入口

使用公开模型 API 基础地址和 Account API Key；不要混用应用 RuntimeSession。具体模型的条件字段由实时服务合同声明。

- 基础地址：/

认证：[{"bearerAuth":[]}]；凭据说明见类型附录。

请求参数

位置	参数	必填	类型	说明
请求头	X-Client-Request-Id	否	string	可选客户端关联值，只能包含可见 ASCII，最多 512 字节；不能替代服务端 requestId。

响应

HTTP 状态	说明	媒体类型与结构
200	请求成功	application/json：OrganizationContext

HTTP 状态	说明	媒体类型与结构
400	请求参数不合法	application/json：PublicError
401	未认证或凭据失效	application/json：PublicError
402	额度不足	application/json：PublicError
403	没有访问权限	application/json：PublicError
404	资源不存在或不可见	application/json：PublicError
409	请求或状态冲突	application/json：PublicError
413	请求体超过大小限制	application/json：PublicError
415	不支持的媒体类型	application/json：PublicError
422	输入无法处理	application/json：PublicError
429	请求过于频繁	application/json：PublicError
500	服务内部错误	application/json：PublicError
502	上游响应异常	application/json：PublicError
503	服务暂不可用	application/json：PublicError
504	上游响应超时	application/json：PublicError

/docs/reference/models/upsertOrganizationPortalUser

同步组织门户用户映射

按门户用户 ID 幂等覆盖映射（显示名、状态、注册时间），供中心按调用 user 标签汇总消费。仅组织当前 active 主账号的 Account API Key 可调用，错误规则同组织查询接口；请求体出现未定义字段返回 400 invalid_argument，中心暂不可写返回 503 portal_user_sync_unavailable（可重试）。

`PUT /v1/organization/portal-users/{portalUserId}`

身份与入口

使用公开模型 API 基础地址和 Account API Key；不要混用应用 RuntimeSession。具体模型的条件字段由实时服务合同声明。

- 基础地址：/

认证：[{"bearerAuth":[]}]；凭据说明见类型附录。

请求参数

位置	参数	必填	类型	说明
路径	portalUserId	是	string	
请求头	X-Client-Request-Id	否	string	可选客户端关联值，只能包含可见 ASCII，最多 512 字节；不能替代服务端 requestId。

请求正文

请求正文必填。

application/json

类型：`OrganizationPortalUserRequest`。完整字段见本文末尾的类型定义附录。

门户用户映射；只接受以下字段，手机号等个人信息不得上送。

字段	类型	必填	说明
displayName	string	是	门户内的账号显示名。；最长长度：128
registeredAt	string	是	门户用户注册时间。；格式：date-time
status	string	是	门户用户状态。；可选值：["active","frozen"]

响应

HTTP 状态	说明	媒体类型与结构
200	请求成功	application/json：OrganizationPortalUser
400	请求参数不合法	application/json：PublicError
401	未认证或凭据失效	application/json：PublicError
402	额度不足	application/json：PublicError
403	没有访问权限	application/json：PublicError
404	资源不存在或不可见	application/json：PublicError

HTTP 状态	说明	媒体类型与结构
409	请求或状态冲突	application/json：PublicError
413	请求体超过大小限制	application/json：PublicError
415	不支持的媒体类型	application/json：PublicError
422	输入无法处理	application/json：PublicError
429	请求过于频繁	application/json：PublicError
500	服务内部错误	application/json：PublicError
502	上游响应异常	application/json：PublicError
503	服务暂不可用	application/json：PublicError
504	上游响应超时	application/json：PublicError

/docs/reference/models/getOrganizationQuota

查询主账号配额与占用

返回主账号在平台侧的同步调用每秒新请求数、异步媒体排队与运行份额的生效上限、来源（account/organization/platform）与当前占用；占用读不到时为 null 并给出 inUseReason。本查询自身按开放查询入口规则准入。身份规则同组织查询接口；规则不可读返回 503 quota_unavailable（可重试）。

GET /v1/organization/quota

身份与入口

使用公开模型 API 基础地址和 Account API Key；不要混用应用 RuntimeSession。具体模型的条件字段由实时服务合同声明。

- 基础地址：/

认证：[{"bearerAuth":[]}]；凭据说明见类型附录。

请求参数

位置	参数	必填	类型	说明
请求头	X-Client-Request-Id	否	string	可选客户端关联值，只能包含可见 ASCII，最多 512 字节；不能替代服务端 requestId。

响应

HTTP 状态	说明	媒体类型与结构
200	请求成功	application/json：OrganizationQuota
400	请求参数不合法	application/json：PublicError
401	未认证或凭据失效	application/json：PublicError
402	额度不足	application/json：PublicError
403	没有访问权限	application/json：PublicError
404	资源不存在或不可见	application/json：PublicError
409	请求或状态冲突	application/json：PublicError
413	请求体超过大小限制	application/json：PublicError
415	不支持的媒体类型	application/json：PublicError
422	输入无法处理	application/json：PublicError
429	请求过于频繁	application/json：PublicError
500	服务内部错误	application/json：PublicError
502	上游响应异常	application/json：PublicError
503	服务暂不可用	application/json：PublicError
504	上游响应超时	application/json：PublicError

查询组织可用服务状态

返回组织可调用服务（按组织集群资格、服务受众与路由池判定）的状态分档、判定依据与近 24 小时 / 7 天可用性；观测陈旧时为 unknown，不显示为健康。身份规则同组织查询接口：仅组织当前 active 主账号的 Account API Key 可调用，非成员 403 organization_membership_required，非主账号 403 organization_primary_required，组织停用 403 organization_access_suspended；路由视图不可读返回 503 service_status_unavailable（可重试）。

`GET /v1/organization/services/status`

身份与入口

使用公开模型 API 基础地址和 Account API Key；不要混用应用 RuntimeSession。具体模型的条件字段由实时服务合同声明。

- 基础地址：`/`

认证：[{"bearerAuth":[]}]；凭据说明见类型附录。

请求参数

位置	参数	必填	类型	说明
请求头	X-Client-Request-Id	否	string	可选客户端关联值，只能包含可见 ASCII，最多 512 字节；不能替代服务端 requestId。
查询	services	否	string	逗号分隔的公开模型别名，最多 50 个；按传入顺序返回，不在可调用集合中的别名返回 maintenance。

响应

HTTP 状态	说明	媒体类型与结构
200	请求成功	application/json：OrganizationServiceStatusList
400	请求参数不合法	application/json：PublicError

HTTP 状态	说明	媒体类型与结构
401	未认证或凭据失效	application/json：PublicError
402	额度不足	application/json：PublicError
403	没有访问权限	application/json：PublicError
404	资源不存在或不可见	application/json：PublicError
409	请求或状态冲突	application/json：PublicError
413	请求体超过大小限制	application/json：PublicError
415	不支持的媒体类型	application/json：PublicError
422	输入无法处理	application/json：PublicError
429	请求过于频繁	application/json：PublicError
500	服务内部错误	application/json：PublicError
502	上游响应异常	application/json：PublicError
503	服务暂不可用	application/json：PublicError
504	上游响应超时	application/json：PublicError

/docs/reference/models/getOrganizationUsageMetrics

查询组织调用体验指标

按服务与时间桶聚合主账号个人工作空间已结束的调用：调用量、成功率、总时延 P50/P95、输出速度与失败分类；首字时延暂无事实来源时为 null。身份规则同组织查询接口；参数错误返回 400 invalid_argument，聚合暂不可用返回 503 usage_metrics_unavailable（可重试）。

GET /v1/organization/usage/metrics

身份与入口

使用公开模型 API 基础地址和 Account API Key；不要混用应用 RuntimeSession。具体模型的条件字段由实时服务合同声明。

- 基础地址：/

认证：[{"bearerAuth":[]}]；凭据说明见类型附录。

请求参数

位置	参数	必填	类型	说明
请求头	X-Client-Request-Id	否	string	可选客户端关联值，只能包含可见 ASCII，最多 512 字节；不能替代服务端 requestId。
查询	from	否	string	窗口下界（含），RFC 3339；默认 to 前 24 小时。
查询	to	否	string	窗口上界（不含），RFC 3339；默认当前时间。窗口最长 31 天。
查询	granularity	否	string	时间桶粒度。
查询	service	否	string	按公开模型别名精确筛选。
查询	user	否	string	按调用时 user 标签精确筛选。
查询	apiKeyId	否	string	按调用所用 Account API Key 标识精确筛选（例如门户专用 Key）；只在主账号自己的调用范围内生效。

响应

HTTP 状态	说明	媒体类型与结构
200	请求成功	application/json: OrganizationUsageMetrics
400	请求参数不合法	application/json：PublicError
401	未认证或凭据失效	application/json：PublicError
402	额度不足	application/json：PublicError
403	没有访问权限	application/json：PublicError
404	资源不存在或不可见	application/json：PublicError
409	请求或状态冲突	application/json：PublicError
413	请求体超过大小限制	application/json：PublicError

HTTP 状态	说明	媒体类型与结构
415	不支持的媒体类型	application/json：PublicError
422	输入无法处理	application/json：PublicError
429	请求过于频繁	application/json：PublicError
500	服务内部错误	application/json：PublicError
502	上游响应异常	application/json：PublicError
503	服务暂不可用	application/json：PublicError
504	上游响应超时	application/json：PublicError

/docs/reference/models/createRerank

重排候选文档

POST /v1/rerank

身份与入口

使用公开模型 API 基础地址和 Account API Key；不要混用应用 RuntimeSession。具体模型的条件字段由实时服务合同声明。

- 基础地址：/

认证：[{"bearerAuth":[]}]；凭据说明见类型附录。

请求参数

位置	参数	必填	类型	说明
请求头	X-Client-Request-Id	否	string	可选客户端关联值，只能包含可见 ASCII，最多 512 字节；不能替代服务端 requestId。

请求正文

请求正文必填。

application/json

类型：`RerankRequest`。完整字段见本文末尾的类型定义附录。

字段	类型	必填	说明
documents	array<string>	是	待重排的候选文本集合，顺序用于与结果 index 对齐。；最少项数：1；最多项数：64
model	string	是	重排模型的公开别名。
query	string	是	用于重排候选内容的查询文本。；最短长度：1；最长长度：1024
top_n	integer	否	仅在模型支持时发送；query 与 documents 总计不超过 65536 个 Unicode 字符。；最小值：1

响应

HTTP 状态	说明	媒体类型与结构
200	请求成功	application/json：RerankResponse
400	请求参数不合法	application/json：PublicError
401	未认证或凭据失效	application/json：PublicError
402	额度不足	application/json：PublicError
403	没有访问权限	application/json：PublicError
404	资源不存在或不可见	application/json：PublicError
409	请求或状态冲突	application/json：PublicError
413	请求体超过大小限制	application/json：PublicError
415	不支持的媒体类型	application/json：PublicError
422	输入无法处理	application/json：PublicError
429	请求过于频繁	application/json：PublicError

HTTP 状态	说明	媒体类型与结构
500	服务内部错误	application/json：PublicError
502	上游响应异常	application/json：PublicError
503	服务暂不可用	application/json：PublicError
504	上游响应超时	application/json：PublicError

/docs/reference/models/createResponse

生成无状态 Responses 回复

无状态 Responses 兼容入口，复用 Chat Completions 的模型合同、鉴权、准入、图片导入和计量。每轮发送完整 input；支持文本、图片、function 工具历史与 typed SSE。store 仅支持 false；不支持服务端会话、previous_response_id、后台任务、内置工具或加密 reasoning 回放。采样和结构化输出能力受所选模型合同约束。

`POST /v1/responses`

身份与入口

使用公开模型 API 基础地址和 Account API Key；不要混用应用 RuntimeSession。具体模型的条件字段由实时服务合同声明。

- 基础地址：/

认证：[{"bearerAuth":[]}]；凭据说明见类型附录。

请求参数

位置	参数	必填	类型	说明
请求头	X-Client-Request-Id	否	string	可选客户端关联值，只能包含可见 ASCII，最多 512 字节；不能替代服务端 requestId。

请求正文

请求正文必填。

application/json

类型：ResponsesRequest。完整字段见本文末尾的类型定义附录。

字段	类型	必填	说明
background	boolean	否	仅支持 false；调用随当前连接执行。；固定值：false
frequency_penalty	number	否	按出现频率调整 token 的采样惩罚。
include	array<string>	否	请求存在时的可选推理加密字段；当前仅返回可兑现内容。
input	string / array<object>	是	字符串，或 message/function_call/function_call_output/明文 reasoning 历史。图片使用 input_image.image_url，支持所选模型允许的 HTTPS/data URI。
instructions	string	否	放在本轮历史前的系统指令。
max_output_tokens	integer	否	本轮最大输出 token 数，受模型预算合同限制。；最小值：1
metadata	object	否	最多 16 项字符串元数据；键和值分别不超过 64 和 512 字符。
model	string	是	所选模型的公开别名。
parallel_tool_calls	boolean	否	是否允许本轮生成多个函数调用。
presence_penalty	number	否	按是否出现调整 token 的采样惩罚。
prompt_cache_key	string	否	可选的提示缓存分区标识；不保证缓存命中。
reasoning	object	否	模型推理配置。
reasoning.effort	string	否	传给所选模型的推理预算级别。；可选值：["none","minimal","low","medium","high","xhigh","max"]
reasoning.summary	string	否	auto 投影模型已有的明文思考内容；不执行独立摘要生成，不生成加密状态。；可选值：["auto"]

字段	类型	必填	说明
seed	integer	否	可选整数随机种子。
service_tier	string	否	仅支持 auto/default，使用模型默认服务级别。；可选值：["auto","default"]
stop	string / array<string>	否	自定义停止字符串，受模型流式及媒体合同约束。
store	boolean	否	仅支持 false；服务端不保存可续接的 Responses 会话。；固定值：false；默认值：false
stream	boolean	否	是否返回带类型的增量 SSE 事件。
temperature	number	否	采样温度，受模型能力约束。
text	object	否	文本输出格式配置。
text.format	object	否	文本或模型支持的结构化输出格式。
tool_choice	string / object	否	自动、禁止、必需或指定名称的函数调用策略。
tools	array<object>	否	函数工具定义；不支持服务器内置工具。
tools[].description	string	否	工具功能说明。
tools[].name	string	是	函数或结构化输出的名称。
tools[].parameters	object	否	函数参数的 JSON Schema，支持范围由模型合同规定。
tools[].strict	boolean	否	是否严格约束函数参数。
tools[].type	string	是	该对象的协议类型。；固定值："function"
top_p	number	否	累积概率采样阈值。
truncation	string	否	仅支持 disabled；超预算明确拒绝。；固定值："disabled"
user	string	否	调用方用户标识。

响应

HTTP 状态	说明	媒体类型与结构
200	请求成功	application/json：ResponsesResponse；text/event-stream：string
400	请求参数不合法	application/json：PublicError
401	未认证或凭据失效	application/json：PublicError
402	额度不足	application/json：PublicError
403	没有访问权限	application/json：PublicError
404	资源不存在或不可见	application/json：PublicError
409	请求或状态冲突	application/json：PublicError
413	请求体超过大小限制	application/json：PublicError
415	不支持的媒体类型	application/json：PublicError
422	输入无法处理	application/json：PublicError
429	请求过于频繁	application/json：PublicError
500	服务内部错误	application/json：PublicError
502	上游响应异常	application/json：PublicError
503	服务暂不可用	application/json：PublicError
504	上游响应超时	application/json：PublicError

/docs/reference/models/listServices

查询服务与动态规格

GET /v1/services

身份与入口

使用公开模型 API 基础地址和 Account API Key；不要混用应用 RuntimeSession。具体模型的条件字段由实时服务合同声明。

- 基础地址：/

认证：[{"bearerAuth":[]}]；凭据说明见类型附录。

请求参数

位置	参数	必填	类型	说明
请求头	X-Client-Request-Id	否	string	可选客户端关联值，只能包含可见 ASCII，最多 512 字节；不能替代服务端 requestId。
查询	serviceKind	否	string	可选，按能力类别精确筛选。

响应

HTTP 状态	说明	媒体类型与结构
200	请求成功	application/json：PublicServiceList
400	请求参数不合法	application/json：PublicError
401	未认证或凭据失效	application/json：PublicError
402	额度不足	application/json：PublicError
403	没有访问权限	application/json：PublicError
404	资源不存在或不可见	application/json：PublicError
409	请求或状态冲突	application/json：PublicError
413	请求体超过大小限制	application/json：PublicError
415	不支持的媒体类型	application/json：PublicError
422	输入无法处理	application/json：PublicError
429	请求过于频繁	application/json：PublicError
500	服务内部错误	application/json：PublicError
502	上游响应异常	application/json：PublicError
503	服务暂不可用	application/json：PublicError
504	上游响应超时	application/json：PublicError

H3 V2：查询视频任务

GET /v2/query/video_generation/{task_id}

身份与入口

使用公开模型 API 基础地址和 Account API Key；不要混用应用 RuntimeSession。具体模型的条件字段由实时服务合同声明。

- 基础地址：/

认证：[{"bearerAuth":[]}]；凭据说明见类型附录。

请求参数

位置	参数	必填	类型	说明
路径	task_id	是	string	

响应

HTTP 状态	说明	媒体类型与结构
200	查询成功；检查 task.status 和 task.error	application/json：MiniMaxV2Task
400	请求失败；保留 request_id，不盲目新建任务。	application/json：MiniMaxV2Error
401	请求失败；保留 request_id，不盲目新建任务。	application/json：MiniMaxV2Error
402	请求失败；保留 request_id，不盲目新建任务。	application/json：MiniMaxV2Error
403	请求失败；保留 request_id，不盲目新建任务。	application/json：MiniMaxV2Error
404	请求失败；保留 request_id，不盲目新建任务。	application/json：MiniMaxV2Error

HTTP 状态	说明	媒体类型与结构
409	请求失败；保留 request_id，不盲目新建任务。	application/json：MiniMaxV2Error
410	请求失败；保留 request_id，不盲目新建任务。	application/json：MiniMaxV2Error
413	请求失败；保留 request_id，不盲目新建任务。	application/json：MiniMaxV2Error
415	请求失败；保留 request_id，不盲目新建任务。	application/json：MiniMaxV2Error
429	请求失败；保留 request_id，不盲目新建任务。	application/json：MiniMaxV2Error
502	请求失败；保留 request_id，不盲目新建任务。	application/json：MiniMaxV2Error
503	请求失败；保留 request_id，不盲目新建任务。	application/json：MiniMaxV2Error
504	请求失败；保留 request_id，不盲目新建任务。	application/json：MiniMaxV2Error

/docs/reference/models/createMiniMaxV2Video

H3 V2：创建视频任务

输出默认以 personal_30d 入库当前用户资源空间，自入库起 30 天到期，可在门户改为永久保存并继续占用配额。顶层 storageMode=temporary_1h 仅后台已对用户或组织授权时允许，输出自入库起 1 小时到期；URL 临时输入保持独立 1 小时策略。同一幂等键不能改变存储模式；缺省与显式 personal_30d 等价，已受理重放沿用原策略。素材地址或来源响应不符合要求返回 400 invalid_source_url；下载失败返回 502 source_fetch_failed，下载超时返回 504 source_fetch_timeout；保留原幂等键重试。

`POST /v2/video_generation`

身份与入口

使用公开模型 API 基础地址和 Account API Key；不要混用应用 RuntimeSession。具体模型的条件字段由实时服务合同声明。

- 基础地址：`/`

认证：[{"bearerAuth":[]}]；凭据说明见类型附录。

请求参数

位置	参数	必填	类型	说明
请求头	Idempotency-Key	否	string	同一业务请求保持Key、正文和此值不变；相同键修改参数返回409。缺省无跨请求幂等保证。

请求正文

请求正文必填。

application/json

类型：`MiniMaxV2Request`。完整字段见本文末尾的类型定义附录。

字段	类型	必填	说明
aigc_watermark	boolean	否	兼容字段，仅支持 false。；固定值：false
content	array<object / object / object / object>	是	恰好一项 text。文生无素材；图生 first_frame；首尾帧 first_frame + last_frame；全能参考角色不得与帧角色混用，至少一个视觉素材。最多9图（有视频8图）、1视频、3音频，总素材12项。；最少项数：1；最多项数：13
duration	integer	是	媒体时长，单位秒。；最小值：5；最大值：15
model	string	是	MiniMax V2 固定兼容模型名 MiniMax-H3，区别于原生 V1 模型别名；大小写须一致。；可选值：["MiniMax-H3"]
optimizePrompt	boolean	否	默认先优化再生成，优化单独计费。false 跳过优化；失败不回退原文生成。；默认值：true
prompt_optimizer	boolean	否	optimizePrompt 的别名；同时出现时必须一致。

字段	类型	必填	说明
ratio	string	是	六档固定画幅：21:9、16:9、4:3、1:1、3:4、9:16；新请求必须显式提供。历史任务可能保留旧值。；可选值：["21:9","16:9","4:3","1:1","3:4","9:16"]
resolution	string	是	V2 清晰度，仅接受大小写精确的 768P。；可选值：["768P"]
storageMode	string	否	平台存储策略，使用顶层字段：默认 personal_30d，将本次主动上传或生成的资源保存到当前用户资源空间，自资源入库起 30 天到期，可在门户“我的资源”改为永久保存，仍占用配额。temporary_1h 自入库起 1 小时到期，仅后台已对所属用户或组织开启临时存储权限时可选；未授权时拒绝，不静默降级。生成请求不改变已引用素材的原到期时间。公网 URL 受控导入的临时输入遵循独立的 1 小时策略，不因本字段进入个人资源空间。；可选值：["personal_30d","temporary_1h"]；默认值："personal_30d"

响应

HTTP 状态	说明	媒体类型与结构
200	已受理，不代表生成成功	application/json：object
400	请求失败；保留 request_id，不盲目新建任务。	application/json：MiniMaxV2Error
401	请求失败；保留 request_id，不盲目新建任务。	application/json：MiniMaxV2Error
402	请求失败；保留 request_id，不盲目新建任务。	application/json：MiniMaxV2Error
403	请求失败；保留 request_id，不盲目新建任务。	application/json：MiniMaxV2Error
404	请求失败；保留 request_id，不盲目新建任务。	application/json：MiniMaxV2Error

HTTP 状态	说明	媒体类型与结构
409	请求失败；保留 request_id，不盲目新建任务。	application/json：MiniMaxV2Error
410	请求失败；保留 request_id，不盲目新建任务。	application/json：MiniMaxV2Error
413	请求失败；保留 request_id，不盲目新建任务。	application/json：MiniMaxV2Error
415	请求失败；保留 request_id，不盲目新建任务。	application/json：MiniMaxV2Error
429	请求失败；保留 request_id，不盲目新建任务。	application/json：MiniMaxV2Error
502	请求失败；保留 request_id，不盲目新建任务。	application/json：MiniMaxV2Error
503	请求失败；保留 request_id，不盲目新建任务。	application/json：MiniMaxV2Error
504	请求失败；保留 request_id，不盲目新建任务。	application/json：MiniMaxV2Error

/docs/reference/models/getMiniMaxV2VideoContent

H3 V2：读取生成视频

GET /v2/video_generation/results/{id}/content

身份与入口

使用公开模型 API 基础地址和 Account API Key；不要混用应用 RuntimeSession。具体模型的条件字段由实时服务合同声明。

- 基础地址：/

认证：[{"miniMaxV2Ticket":[]}]；凭据说明见类型附录。

请求参数

位置	参数	必填	类型	说明
路径	id	是	string	
请求头	Range	否	string	

响应

HTTP 状态	说明	媒体类型与结构
200	MP4 内容；HEAD 仅返回响应头	video/mp4：string
206	指定字节范围的 MP4 内容	
400	请求失败；保留 request_id，不盲目新建任务。	application/json：MiniMaxV2Error
401	请求失败；保留 request_id，不盲目新建任务。	application/json：MiniMaxV2Error
402	请求失败；保留 request_id，不盲目新建任务。	application/json：MiniMaxV2Error
403	请求失败；保留 request_id，不盲目新建任务。	application/json：MiniMaxV2Error
404	请求失败；保留 request_id，不盲目新建任务。	application/json：MiniMaxV2Error
409	请求失败；保留 request_id，不盲目新建任务。	application/json：MiniMaxV2Error
410	请求失败；保留 request_id，不盲目新建任务。	application/json：MiniMaxV2Error
413	请求失败；保留 request_id，不盲目新建任务。	application/json：MiniMaxV2Error
415	请求失败；保留 request_id，不盲目新建任务。	application/json：MiniMaxV2Error
416	不可满足的字节范围	
429	请求失败；保留 request_id，不盲目新建任务。	application/json：MiniMaxV2Error
502	请求失败；保留 request_id，不盲目新建任务。	application/json：MiniMaxV2Error

HTTP 状态	说明	媒体类型与结构
503	请求失败；保留 request_id，不盲目新建任务。	application/json：MiniMaxV2Error
504	请求失败；保留 request_id，不盲目新建任务。	application/json：MiniMaxV2Error

/docs/reference/models/headMiniMaxV2VideoContent

H3 V2：读取生成视频

HEAD /v2/video_generation/results/{id}/content

身份与入口

使用公开模型 API 基础地址和 Account API Key；不要混用应用 RuntimeSession。具体模型的条件字段由实时服务合同声明。

- 基础地址：/

认证：[{"miniMaxV2Ticket":[]}]；凭据说明见类型附录。

请求参数

位置	参数	必填	类型	说明
路径	id	是	string	
请求头	Range	否	string	

响应

HTTP 状态	说明	媒体类型与结构
200	MP4 内容；HEAD 仅返回响应头	video/mp4：string
206	指定字节范围的 MP4 内容	
400	请求失败；保留 request_id，不盲目新建任务。	application/json：MiniMaxV2Error

HTTP 状态	说明	媒体类型与结构
401	请求失败；保留 request_id，不盲目新建任务。	application/json：MiniMaxV2Error
402	请求失败；保留 request_id，不盲目新建任务。	application/json：MiniMaxV2Error
403	请求失败；保留 request_id，不盲目新建任务。	application/json：MiniMaxV2Error
404	请求失败；保留 request_id，不盲目新建任务。	application/json：MiniMaxV2Error
409	请求失败；保留 request_id，不盲目新建任务。	application/json：MiniMaxV2Error
410	请求失败；保留 request_id，不盲目新建任务。	application/json：MiniMaxV2Error
413	请求失败；保留 request_id，不盲目新建任务。	application/json：MiniMaxV2Error
415	请求失败；保留 request_id，不盲目新建任务。	application/json：MiniMaxV2Error
416	不可满足的字节范围	
429	请求失败；保留 request_id，不盲目新建任务。	application/json：MiniMaxV2Error
502	请求失败；保留 request_id，不盲目新建任务。	application/json：MiniMaxV2Error
503	请求失败；保留 request_id，不盲目新建任务。	application/json：MiniMaxV2Error
504	请求失败；保留 request_id，不盲目新建任务。	application/json：MiniMaxV2Error

/docs/reference/models/schema-appendix

模型 API 类型定义附录

所有接口使用以下同源类型定义；字段标识、枚举值和单位与在线协议一致。

AdmissionPreviewRequest

类型：`object`

字段	类型	必填	说明
model	string	是	支持异步准入的服务公开模型别名。
request	object / null	否	待预检的创建请求体；省略或为 null 时视为空对象。平台使用所选别名覆盖 request.model。


```
{
  "model": "minimax-h3",
  "request": {
    "generationMode": "t2v",
    "model": "minimax-h3",
    "orientation": "landscape",
    "prompt": "雨夜站台，镜头缓慢推进，人物自然递伞后共同离开。",
    "resolutionTier": "768p",
    "seconds": 5,
    "seed": 1234,
    "storageMode": "personal_30d"
  }
}
```


AdmissionPreviewResponse

类型：`object`

字段	类型	必填	说明
admission	object	是	当前准入判断及可选重试建议；不预占资源，也不保证后续请求被受理。
generatedAt	string	是	当前快照生成时间。；格式：date-time
publicationId	string	是	所选服务的关联标识，仅供查看；提交请求仍使用模型别名。
serviceKind	string	是	能力类别，例如 video_generation。

AudioGenerationRequest

类型：`object`

字段	类型	必填	说明
generationMode	string	否	所选服务支持的生成模式。
input_video_asset_id	string	否	已上传的视频素材，用于视频音效；纯文本语音合成不传此字段。
input_video_url	string	否	可选：与 input_video_asset_id 并列的外部来源，二选一。只接受公网 HTTPS（443 端口）；平台在接纳前经受控下载器摄入为当前调用身份和模型的一小时临时素材，再按资源 ID 原链路处理，等同先调用 /v1/assets/input。同槽位同时传 ID 与 URL 返回 400 ambiguous_input_source；同请求重复 URL 只下载一次；来源被拒为 400 invalid_source_url，下载失败 502 source_fetch_failed，超时 504 source_fetch_timeout；部署关闭本能力时返回 400 input_url_not_enabled。
instruct	string	否	voice_design 模式必填；Qwen3-TTS 最多 160 个字符。
model	string	是	音频服务公开模型别名；语音、音乐与视频音效使用不同的条件字段。
prompt	string	否	服务支持时填写音效或音乐描述。
seconds	integer	否	请求时长，单位秒；仅适用于声明支持该字段的模型。
seed	integer	否	可选随机种子；幂等重试时保持原值。
speaker	string	否	custom_voice 模式必填；使用服务已公布的音色名称。

字段	类型	必填	说明
storageMode	string	否	平台存储策略，使用顶层字段：默认 personal_30d，将本次主动上传或生成的资源保存到当前用户资源空间，自资源入库起 30 天到期，可在门户“我的资源”改为永久保存，仍占用配额。temporary_1h 自入库起 1 小时到期，仅后台已对所属用户或组织开启临时存储权限时可选；未授权时拒绝，不静默降级。生成请求不改变已引用素材的原到期时间。公网 URL 受控导入的临时输入遵循独立的 1 小时策略，不因本字段进入个人资源空间。；可选值：["personal_30d","temporary_1h"]；默认值："personal_30d"
text	string	否	待合成的文本；Qwen3-TTS 的字符上限由服务声明，其他模型使用各自限制。；最长长度：600
voiceMode	string	否	Qwen3-TTS 模式：text（默认普通朗读）、custom_voice（内置音色）、voice_design（音色描述）。

AudioTranscriptionRequest

类型：`object`

字段	类型	必填	说明
file	string	是	一个非空文件，不超过 25 MiB。MIME 支持 audio/wav、x-wav、mpeg、mp3、mp4、x-m4a、webm、ogg、flac；octet-stream 须有受支持的文件扩展名。；格式：binary
language	string	否	仅在所选模型明确支持时传入。
model	string	是	语音转写服务的公开模型别名。

字段	类型	必填	说明
response_format	string	否	仅在所选模型支持时使用；通常返回 JSON。

AudioTranscriptionResponse

类型：`object`

字段	类型	必填	说明
duration	number	否	媒体时长，单位秒。
language	string	否	语言代码；合法值由所选服务声明。
segments	array<object>	否	按顺序排列的转写片段。
segments[].end	number	否	该片段结束时间，单位秒。
segments[].id	integer	否	当前对象的唯一标识。
segments[].start	number	否	该片段开始时间，单位秒。
segments[].text	string	否	文本内容。
text	string	否	文本内容。

AudioTranscriptionURLRequest

类型：`object`

字段	类型	必填	说明
input_audio_url	string	是	无需登录即可下载的公网 HTTPS 音频地址，可使用有效期内的签名 URL；与 file 互斥。平台受限下载并验证实际格式、大小和时长。；匹配规则：^https://；格式：uri
language	string	否	仅在所选模型明确支持时传入。
model	string	是	语音转写服务的公开模型别名。
response_format	string	否	仅在所选模型支持时使用；通常返回 JSON。

BillingBalance

主账号个人账户积分余额。

类型：`object`

字段	类型	必填	说明
availableMilliPoints	integer	是	可用余额（千分之一积分）。；格式：int64
object	string	是	对象类型。；固定值："billing.balance"
observedAt	string	是	余额观测时间。；格式：date-time
reservedMilliPoints	integer	是	调用中冻结的金额（千分之一积分）。；格式：int64

BillingCall

一次调用 attempt 的计费明细；同一 callId 可能因结算、失败或退回再次出现，按 callId 幂等覆盖。

类型：`object`

字段	类型	必填	说明
amountMilliPoints	integer / null	是	最终扣费金额（千分之一积分，已扣除退回）；final=false 时为 null。；格式：int64
apiKeyId	string / null	否	调用所用 Account API Key 标识；平台旧版本未提供时为 null。
billingAccount	object / null	否	调用计费的账号工作空间（主账号变更后可能是前任主账号）；未知时为 null。
billingAccount.current	boolean	是	是否为调用方 Key 当前所属的工作空间。
billingAccount.ref	string	是	脱敏的计费工作空间标识，前后各 4 位。
callId	string	是	调用 attempt 的稳定标识，幂等覆盖键。
createdAt	string	是	调用开始时间。；格式：date-time

字段	类型	必填	说明
final	boolean	是	金额是否已最终确定；false 时不得按零结算。
model	string	是	调用时使用的公开模型别名。
requestId	string	是	平台请求标识，与推理响应头 X-Request-Id 对应。
serviceKind	string	是	服务能力类别。
status	string	是	计费状态：pending 尚未结算；settled 已扣费；failed 未扣费结束；refunded 已部分或全部退回。；可选值: ["pending","settled","failed","refunded"]
updatedAt	string	是	最近一次状态或金额变化时间，游标排序键。；格式: date-time
usage	array<object>	是	已观测到的计量项；尚未计量时为空数组。
usage[].components	array<object>	否	组合计价的分项；单一计量项省略。
usage[].components[].code	string	是	分项代码。
usage[].components[].meter	string	是	分项计量项。
usage[].components[].quantity	integer	是	分项数量。；格式：int64
usage[].components[].unit	string	是	分项计量单位。
usage[].meter	string	是	计量项，例如 input_tokens、output_tokens 或媒体资源类型。
usage[].quantity	integer	是	该计量项的数量。；格式: int64
usage[].spec	string	否	计价所用规格码；默认规格省略。
usage[].unit	string	是	计量单位，例如 token、video_second。
user	string / null	是	调用时 OpenAI 兼容 user 字段的规范化值；未提供时为 null。

BillingCallList

按 (updatedAt, callId) 升序的调用明细页。

类型：`object`

字段	类型	必填	说明
data	array<BillingCall>	是	本页调用明细。
hasMore	boolean	是	是否还有更多已变化的记录可立即读取。
nextCursor	string / null	是	下一次增量拉取使用的不透明游标；末页游标会保留最近数秒的重叠回读窗口。
object	string	是	对象类型。；固定值："list"

CatalogVoice

类型：`object`

字段	类型	必填	说明
description	string	否	面向调用方的说明。
displayName	string	是	供用户阅读的显示名称。
previewAssetId	string	否	可选预览素材标识。
primaryLanguage	string	是	主要支持的语言。
supportedLanguages	array<string>	是	支持的语言代码集合。
tags	array<string>	否	公开分类标签。
version	integer	是	当前对象的版本。
voiceId	string	是	音色目录返回的公开音色标识。

CatalogVoiceList

类型：`object`

字段	类型	必填	说明
data	array<CatalogVoice>	是	本操作的数据集合；不同接口不保证共用同一种信封。
object	string	是	返回对象的类型标记。；固定值："list"

ChatRequest

类型：`object`

字段	类型	必填	说明
chat_template_kwargs	object	否	Qwen3.8 兼容的模板参数；建议使用顶层字段。同名字段不能与顶层冲突；json_schema 要求两处均关闭思考。
chat_template_kwargs.enable_thinking	boolean	否	仅支持思考的 Qwen3.8 服务：true 开启；与 reasoning_effort 同时传入时必须一致。默认关闭。；默认值：false
chat_template_kwargs.preserve_thinking	boolean	否	Qwen3.8 模板是否保留已有 assistant 思考历史；不是缓存或服务端会话开关。；默认值：true
chat_template_kwargs.reasoning_effort	string	否	Qwen3.8：none 关闭，minimal/low 映射 low，medium 映射 medium，high/xhigh/max 映射 xhigh；非 none 会开启思考。仅 enable_thinking=true 且省略此字段时采用 xhigh。不单独扩大输出预算。；可选值：["none","minimal","low","medium","high","xhigh","max"]
enable_thinking	boolean	否	仅支持思考的 Qwen3.8 服务：true 开启；与 reasoning_effort 同时传入时必须一致。默认关闭。；默认值：false

字段	类型	必填	说明
max_tokens	integer	否	本次输出预算，受模型输入输出合计上限约束；默认值由所选服务的冻结合同声明，包含思考与最终回答。思考请求应显式设置，预算耗尽返回 finish_reason=length，可能没有最终答案。；最小值：1
messages	array<object>	是	按对话顺序排列的消息集合。
messages[].content	string / array<object>	是	消息内容；支持字符串或模型允许的多模态片段。
messages[].reasoning_content	string / null	否	保留原始 assistant 消息中已有的思考历史，不拼入 user 正文。
messages[].role	string	是	消息或素材的角色，枚举含义依所属对象确定。；可选值：["system","user","assistant","tool"]
model	string	是	文本或视觉对话服务的公开模型别名。
preserve_thinking	boolean	否	Qwen3.8 模板是否保留已有 assistant 思考历史；不是缓存或服务端会话开关。；默认值：true
reasoning_effort	string	否	Qwen3.8：none 关闭，minimal/low 映射 low，medium 映射 medium，high/xhigh/max 映射 xhigh；非 none 会开启思考。仅 enable_thinking=true 且省略此字段时采用 xhigh。不单独扩大输出预算。；可选值：["none","minimal","low","medium","high","xhigh","max"]
stream	boolean	否	仅在服务支持流式输出时返回 SSE。采样参数须由模型合同允许。；默认值：false
stream_options	object	否	流式输出选项，仅在 stream=true 时使用。
stream_options.include_usage	boolean	否	流式结束前返回实际用量；该事件可能没有 choices。；默认值：false

ChatResponse

类型：`object`

字段	类型	必填	说明
choices	array<object>	否	模型返回的候选回复集合。
choices[].finish_reason	string / null	否	停止原因；length 表示达到输出上限，不能当作内容完整。
choices[].index	integer	否	结果对应的输入或候选位置，从 0 开始。
choices[].message	object	否	可读说明或消息对象，按所属响应结构解析。
choices[].message.content	string / null	否	消息内容；支持字符串或模型允许的多模态片段。
choices[].message.reasoning_content	string / null	否	Qwen3.8 开启思考时返回的思考内容，与 content 最终答案分别处理；关闭时可省略。
choices[].message.role	string	否	消息或素材的角色，枚举含义依所属对象确定。
choices[].message.tool_calls	array<object>	否	模型请求的工具调用集合，业务方应按权限验证后执行。
choices[].message.tool_calls[].function	object	否	工具函数调用对象。
choices[].message.tool_calls[].function.arguments	string	否	工具调用参数，通常为 JSON 字符串；执行前须按工具定义验证。
choices[].message.tool_calls[].function.name	string	否	当前对象或工具的名称。
choices[].message.tool_calls[].id	string	否	当前对象的唯一标识。
choices[].message.tool_calls[].type	string	否	当前对象、消息片段或工具的类型。
created	integer	否	创建时间，通常为 Unix 秒时间戳，以所属对象合同为准。
id	string	否	当前对象的唯一标识。
model	string	否	服务发现返回的公开模型别名。

字段	类型	必填	说明
object	string	否	返回对象的类型标记。
usage	TokenUsage	否	类型定义：TokenUsage

DocumentParseJob

类型：`object`

字段	类型	必填	说明
assets	array<MediaResultAsset>	否	兼容结果集合；新客户端优先读取 outputs。
completedAt	string	否	任务完成时间，采用带时区的日期时间。；格式：date-time
createdAt	string	否	创建时间，采用带时区的日期时间。；格式：date-time
elapsedSeconds	number	否	任务墙钟总耗时，单位秒；不是计费生成时长。；最小值：0
errorCode	string	否	失败任务的机器错误码；按原因修正输入或恢复原任务，不重复创建未知结果任务。
errorMessage	string	否	可安全展示的失败说明。
id	string	否	jobId 的兼容别名。
jobId	string	是	受理后持久保存的任务 ID；用 GET /v1/jobs/{id}?model=... 查询原任务。
jobUrl	string	是	带鉴权的绝对轮询地址；携带原 API Key 原样请求，不用 publicationId 自行拼接。；格式：uri
manifest	object	否	成功后的 document-parse-result/v1 清单；artifacts 包含 markdown、content_list、middle_json 和可选 images_archive 的 role、contentType、sizeBytes 与 sha256，通过摘要关联 outputs。
model	string	否	公开模型别名。

字段	类型	必填	说明
object	string	否	返回对象的类型标记。；固定值："document.parse.job"
outputs	array<MediaResultAsset>	否	成功后的权威结果集合；优先于兼容结果字段使用。
queuedSeconds	number	否	排队耗时，单位秒。；最小值：0
refreshAfterSeconds	integer	否	建议的下次刷新间隔，单位秒。；最小值：0
requestId	string	是	服务端生成的请求追踪标识。
retryAfterSeconds	integer	否	建议重试或轮询前等待的秒数。；最小值：0
runningSeconds	number	否	执行耗时，单位秒。；最小值：0
serviceKind	string	否	能力类别，例如 video_generation。；固定值："document_parse"
stage	string	否	任务当前阶段；HTTP 200 不代表生成成功。
stageMessage	string	否	可读的当前阶段说明，不用于驱动状态分支。
startedAt	string	否	任务开始执行时间。；格式：date-time
status	string	是	文档异步任务状态；不开放用户取消。；可选值：["queued","running","succeeded","failed"]
statusCode	integer	否	关联的 HTTP 状态码。
usage	object	否	可信 sourcePages、processedPages 和页计量事实；pages 的预占值不等同最终实际用量。

```
{
  "id": "job_document_example",
  "jobId": "job_document_example",
  "jobUrl": "https://api.example.test/v1/jobs/job_document_example?model=mineru-document-parse",
  "model": "mineru-document-parse",
  "object": "document.parse.job",
  "requestId": "req_example",
  "serviceKind": "document_parse",
  "stage": "queued",
  "status": "queued"
}
```


允许扩展属性；结构：object

DocumentParseRequest

类型：`object`

字段	类型	必填	说明
input_document_asset_id	string	否	当前调用身份和模型范围内已上传的文档素材。
input_document_url	string	否	可选：与 input_document_asset_id 并列的外部来源，二选一。只接受公网 HTTPS（443 端口）；平台在接纳前经受控下载器摄入为当前调用身份和模型的一小时临时素材，再按资源 ID 原链路处理，等同先调用 /v1/assets/input。同槽位同时传 ID 与 URL 返回 400 ambiguous_input_source；同请求重复 URL 只下载一次；来源被拒为 400 invalid_source_url，下载失败 502 source_fetch_failed，超时 504 source_fetch_timeout；部署关闭本能力时返回 400 input_url_not_enabled。
language	string	否	语言代码；合法值由所选服务声明。；可选值：["ch","en"]
model	string	是	服务发现返回的公开模型别名。
original_name	string	否	原始文件名，不包含目录路径。

字段	类型	必填	说明
page_range	object	否	待处理的闭区间页码范围。
page_range.end_page	integer	是	结束页，闭区间；页码基准以所选解析服务为准。；最小值：1；最大值：500
page_range.start_page	integer	是	仅 PDF 接受页范围，从 1 开始的闭区间起点；Office 和图片不接受 page_range。；最小值：1；最大值：500
pages	integer	否	预占页数上限，范围 1–500；PDF 全文及 DOCX/PPTX/XLSX 必填，图片按一页。最终按实际 processedPages 结算。；最小值：1；最大值：500
preset	string	否	所选服务支持的解析预设。；可选值：["hybrid-medium","hybrid-high"]
storageMode	string	否	平台存储策略，使用顶层字段：默认 personal_30d，将本次主动上传或生成的资源保存到当前用户资源空间，自资源入库起 30 天到期，可在门户“我的资源”改为永久保存，仍占用配额。temporary_1h 自入库起 1 小时到期，仅后台已对所属用户或组织开启临时存储权限时可选；未授权时拒绝，不静默降级。生成请求不改变已引用素材的原到期时间。公网 URL 受控导入的临时输入遵循独立的 1 小时策略，不因本字段进入个人资源空间。；可选值：["personal_30d","temporary_1h"]；默认值："personal_30d"

EffectiveMediaSpec

平台归一化后的实际模型规格；请求秒数、原生帧数与产物时长是不同概念。

类型：`object`

字段	类型	必填	说明
aspectRatio	string	否	实际生效的画幅比例。

字段	类型	必填	说明
contractVersion	string	否	实际生效的媒体合同版本。
fps	number	否	每秒帧数；只在模型明确开放时可配置。
generationMode	string	否	归一化后的生成模式。
height	integer	否	高度，单位像素。
orientation	string	否	画幅方向；landscape 为横向，portrait 为纵向。
outputSizeCode	string	否	实际生效的固定输出尺寸代码。
resolutionTier	string	否	归一化后的清晰度档位。
seconds	integer	否	请求时长，单位秒；仅适用于声明支持该字段的模型。
width	integer	否	宽度，单位像素。

允许扩展属性；结构：object

EmbeddingRequest

类型：`object`

字段	类型	必填	说明
input	string / array<string>	是	去除首尾空白后不可为空；最多 64 个非空文本，每项最多 8192 个 Unicode 字符。
model	string	是	向量模型的公开别名。

EmbeddingResponse

类型：`object`

字段	类型	必填	说明
data	array<object>	否	本操作的数据集合；不同接口不保证共用同一种信封。
data[].embedding	array<number>	否	与输入对应的向量数组；维度由模型决定。

字段	类型	必填	说明
data[].index	integer	否	结果对应的输入或候选位置，从 0 开始。
data[].object	string	否	返回对象的类型标记。
model	string	否	服务发现返回的公开模型别名。
object	string	否	返回对象的类型标记。
usage	TokenUsage	否	类型定义：TokenUsage

H3GenerationReference

类型：`object`

字段	类型	必填	说明
assetId	string	否	已上传的输入素材 ID；不是 URL 或本地文件路径。
mediaType	string	否	省略时由 role 推导；显式传入时必须与角色对应的类型一致。；可选值: ["image","video","audio"]
role	string	是	消息或素材的角色，枚举含义依所属对象确定。；可选值: ["reference_image","reference_video","reference_audio"]
sha256	string	否	可选完整性提示；平台会自行校验并固定素材摘要。
url	string	否	可选：行内未给 assetId 时的外部来源（公网 HTTPS，图片另可为 ），平台在接纳前摄入为一小时临时素材并改写为 assetId。行内已有 assetId 时忽略 url，不下载。来源被拒为 400 invalid_source_url；部署关闭时返回 400 input_url_not_enabled。

ImageGenerationRequest

类型：`object`

字段	类型	必填	说明
aspectRatio	string	否	受控输出长宽比；H3 由服务能力合同返回，例如 21:9、16:9、4:3、1:1、3:4、9:16。
generationMode	string	否	模型支持的生成模式，例如 t2i（文生图）、i2i（图生图）。
imageCount	integer	否	请求生成图片数量；以模型允许范围为准。；最小值：1
input_image_asset_id	string	否	受控单张图片素材，仅用于支持图片输入的模型。
input_image_asset_ids	array<string>	否	按顺序排列的输入图片 ID；仅支持多图的模型可用，不与单图字段混传。
input_image_url	string	否	可选：与 input_image_asset_id 并列的外部来源，二选一。只接受公网 HTTPS（443 端口），另接受 图片数据；平台在接纳前经受控下载器摄入为当前调用身份和模型的一小时临时素材，再按资源 ID 原链路处理，等同先调用 /v1/assets/input。同槽位同时传 ID 与 URL 返回 400 ambiguous_input_source；同请求重复 URL 只下载一次；来源被拒为 400 invalid_source_url，下载失败 502 source_fetch_failed，超时 504 source_fetch_timeout；部署关闭本能力时返回 400 input_url_not_enabled。

字段	类型	必填	说明
input_image_urls	array<string>	否	有序 URL 数组，按位置改写为 input_image_asset_ids。可选：与 input_image_asset_ids 并列的外部来源，二选一。只接受公网 HTTPS（443 端口），另接受 图片数据；平台在接纳前经受控下载器摄入为当前调用身份和模型的一小时临时素材，再按资源 ID 原链路处理，等同先调用 /v1/assets/input。同槽位同时传 ID 与 URL 返回 400 ambiguous_input_source；同请求重复 URL 只下载一次；来源被拒为 400 invalid_source_url，下载失败 502 source_fetch_failed，超时 504 source_fetch_timeout；部署关闭本能力时返回 400 input_url_not_enabled。
model	string	是	图片模型的公开别名。
orientation	string	否	画幅方向，合法值按模型规格选择。
outputSizeCode	string	否	模型公布的输出尺寸代码；不要自行拼接尺寸。
prompt	string	是	图片生成提示词；字符上限由所选服务声明。
qualityTier	string	否	模型公布的质量档位。

字段	类型	必填	说明
reference_image_url	string	否	可选：与 reference_image_asset_id 并列的外部来源，二选一。只接受公网 HTTPS（443 端口），另接受 图片数据；平台在接纳前经受控下载器摄入为当前调用身份和模型的一小时临时素材，再按资源 ID 原链路处理，等同先调用 /v1/assets/input。同槽位同时传 ID 与 URL 返回 400 ambiguous_input_source；同请求重复 URL 只下载一次；来源被拒为 400 invalid_source_url，下载失败 502 source_fetch_failed，超时 504 source_fetch_timeout；部署关闭本能力时返回 400 input_url_not_enabled。
reference_image_urls	array<string>	否	有序 URL 数组，按位置改写为 reference_image_asset_ids。可选：与 reference_image_asset_ids 并列的外部来源，二选一。只接受公网 HTTPS（443 端口），另接受 图片数据；平台在接纳前经受控下载器摄入为当前调用身份和模型的一小时临时素材，再按资源 ID 原链路处理，等同先调用 /v1/assets/input。同槽位同时传 ID 与 URL 返回 400 ambiguous_input_source；同请求重复 URL 只下载一次；来源被拒为 400 invalid_source_url，下载失败 502 source_fetch_failed，超时 504 source_fetch_timeout；部署关闭本能力时返回 400 input_url_not_enabled。
resolutionTier	string	否	所选模型支持的清晰度档位。
seed	integer	否	可选随机种子；幂等重试时保持原值。

字段	类型	必填	说明
size	string	否	图片尺寸，格式为宽x高，例如1024x1024；仅可使用所选服务 mediaSpecContract 声明的尺寸。
storageMode	string	否	平台存储策略，使用顶层字段：默认 personal_30d，将本次主动上传或生成的资源保存到当前用户资源空间，自资源入库起 30 天到期，可在门户“我的资源”改为永久保存，仍占用配额。temporary_1h 自入库起 1 小时到期，仅后台已对所属用户或组织开启临时存储权限时可选；未授权时拒绝，不静默降级。生成请求不改变已引用素材的原到期时间。公网 URL 受控导入的临时输入遵循独立的 1 小时策略，不因本字段进入个人资源空间。；可选值：["personal_30d","temporary_1h"]；默认值："personal_30d"

InputAssetUploadResponse

类型：`object`

字段	类型	必填	说明
asset	object	是	本次上传返回的素材对象。
asset.assetId	string	是	素材唯一标识，不是任务 ID。
asset.byteSize	integer	是	文件大小，单位字节。；格式：int64
asset.createdAt	string	是	创建时间，采用带时区的日期时间。；格式：date-time
asset.expiresAt	string	是	本次返回的资源到期时间，以服务端为准；读取和引用资源不续期。到期后停止用户访问，后台异步清理。；格式：date-time
asset.ingestedAt	string	是	输入资源首次入库时间；复用同一资源不重新起算保留期。；格式：date-time

字段	类型	必填	说明
asset.kind	string	是	对象或媒体类别。；可选值: ["image","audio","video","document"]
asset.lastUploadedAt	string	是	最近上传时间。；格式：date-time
asset.mimeType	string	是	实际媒体 MIME 类型，例如 video/mp4。
asset.publicationId	string	是	服务关联标识；调用时继续使用公开模型别名。
asset.reused	boolean	是	是否复用了相同身份和模型范围内的素材。
asset.status	string	是	对象当前状态，必须按所属操作的状态集合判断。；固定值："available"
asset.storageBackend	string	否	存储类型说明，不用于自行拼接存储地址。
asset.storageMode	string	是	该资源实际采用的存储模式；历史资源可能未记录，不应据此补造 30 天期限。永久保存及实际到期时间以资源当前保留策略和 expiresAt 为准。；可选值: ["personal_30d","temporary_1h"]
usage	object	否	本次调用用量；不等同于最终价格。

MediaGenerationRequest

公开视频接口使用顶层字段，不使用正式应用的 params/mediaSpec 包装。H3 必须提供 generationMode、resolutionTier、orientation、seconds，支持 544p/768p、六种受控画布和 5 ~ 15 秒，提示词最多 7000 字符。t2v 不传素材；i2v 传首图；首尾帧传两张图；全能参考传分类素材。其他模型的条件必填和合法组合由 mediaSpecContract 声明；不要向 H3 传 fps、frames、width、height、steps 或运行时设置。

类型：`object`

字段	类型	必填	说明
aspectRatio	string	否	H3 受控画面比例，例如 21:9、16:9、4:3、1:1、3:4、9:16；需与 outputSizeCode 配套。

字段	类型	必填	说明
characterMode	string	否	Wan2.2 Animate 专用：动作迁移或角色替换；需要人物图和驱动视频，不传 prompt。其他视频模型不使用本字段。；可选值：["motion_transfer","character_replace"]
enableRife	boolean	否	仅部分普通 Wan 视频服务支持的插帧选项；H3、Wan Animate 和 X-Dub 不接受。须先确认所选服务能力。
end_image_asset_id	string	否	已上传的尾帧图片；仅用于 H3 首尾帧等明确支持尾帧的模式。
end_image_url	string	否	可选：与 end_image_asset_id 并列的外部来源，二选一。只接受公网 HTTPS（443 端口），另接受 图片数据；平台在接纳前经受控下载器摄入为当前调用身份和模型的一小时临时素材，再按资源 ID 原链路处理，等同先调用 /v1/assets/input。同槽位同时传 ID 与 URL 返回 400 ambiguous_input_source；同请求重复 URL 只下载一次；来源被拒为 400 invalid_source_url，下载失败 502 source_fetch_failed，超时 504 source_fetch_timeout；部署关闭本能力时返回 400 input_url_not_enabled。
generationMode	string	否	H3 必填，可选 t2v、i2v、first_last_frame_video、universal_reference_video；其他模型使用各自公布的模式。
input_audio_asset_id	string	否	明确支持音频输入的模型所需素材；不是 H3 全能参考的独立顶层字段。

字段	类型	必填	说明
input_audio_url	string	否	可选：与 input_audio_asset_id 并列的外部来源，二选一。只接受公网 HTTPS（443 端口）；平台在接纳前经受控下载器摄入为当前调用身份和模型的一小时临时素材，再按资源 ID 原链路处理，等同先调用 /v1/assets/input。同槽位同时传 ID 与 URL 返回 400 ambiguous_input_source；同请求重复 URL 只下载一次；来源被拒为 400 invalid_source_url，下载失败 502 source_fetch_failed，超时 504 source_fetch_timeout；部署关闭本能力时返回 400 input_url_not_enabled。
input_image_asset_id	string	否	已上传的输入图片；H3 用于图生和首尾帧，Wan Animate 用于人物图；其他模型按服务合同使用。
input_image_asset_ids	array<string>	否	有序输入图片 ID；Bernini r2v 使用 1 ~ 5 张，rv2v 使用恰好 1 张。不是 H3 全能参考字段；仅按当前模型能力使用。
input_image_url	string	否	可选：与 input_image_asset_id 并列的外部来源，二选一。只接受公网 HTTPS（443 端口），另接受 图片数据；平台在接纳前经受控下载器摄入为当前调用身份和模型的一小时临时素材，再按资源 ID 原链路处理，等同先调用 /v1/assets/input。同槽位同时传 ID 与 URL 返回 400 ambiguous_input_source；同请求重复 URL 只下载一次；来源被拒为 400 invalid_source_url，下载失败 502 source_fetch_failed，超时 504 source_fetch_timeout；部署关闭本能力时返回 400 input_url_not_enabled。

字段	类型	必填	说明
input_image_urls	array<string>	否	有序 URL 数组，按位置改写为 input_image_asset_ids。可选：与 input_image_asset_ids 并列的外部来源，二选一。只接受公网 HTTPS（443 端口），另接受 图片数据；平台在接纳前经受控下载器摄入为当前调用身份和模型的一小时临时素材，再按资源 ID 原链路处理，等同先调用 /v1/assets/input。同槽位同时传 ID 与 URL 返回 400 ambiguous_input_source；同请求重复 URL 只下载一次；来源被拒为 400 invalid_source_url，下载失败 502 source_fetch_failed，超时 504 source_fetch_timeout；部署关闭本能力时返回 400 input_url_not_enabled。
input_video_asset_id	string	否	明确支持视频输入的模型所需素材；不是 H3 全能参考的独立顶层字段。
input_video_url	string	否	可选：与 input_video_asset_id 并列的外部来源，二选一。只接受公网 HTTPS（443 端口）；平台在接纳前经受控下载器摄入为当前调用身份和模型的一小时临时素材，再按资源 ID 原链路处理，等同先调用 /v1/assets/input。同槽位同时传 ID 与 URL 返回 400 ambiguous_input_source；同请求重复 URL 只下载一次；来源被拒为 400 invalid_source_url，下载失败 502 source_fetch_failed，超时 504 source_fetch_timeout；部署关闭本能力时返回 400 input_url_not_enabled。
model	string	是	GET /v1/services 返回的公开模型别名；不能使用部署名或工作负载名。

字段	类型	必填	说明
optimizePrompt	boolean	否	仅公开 Account API Key H3 四模式生成缺省开启。部署未激活时，缺省或 true 返回 503 prompt_optimization_capability_unavailable；查询 automaticPromptOptimization.available 获取编排可用性，它不代表模型实时健康。显式 false 完全跳过优化且不收优化费；已手动优化的 prompt 应传 false。其他模型不会自动优化，true 返回参数错误。后台准备最多两次常规优化，每次完整超长候选最多压缩一次，最多四次模型调用按实际 token 单独计费。两次失败则 Job 返回 prompt_optimization_failed，不回退原 prompt；未知结果恢复原 attempt，不新发模型调用。
orientation	string	否	画幅方向，合法值按模型规格选择。；可选值: ["landscape","portrait"]
outputSizeCode	string	否	H3 受控画布代码；服务能力合同返回的六种 544p 或六种 768p 代码之一。
prompt	string	否	生成提示词；H3 去除首尾空白后为 1 ~ 7000 个 Unicode 字符。Wan Animate、X-Dub 等无提示词模式不传。
referenceInputs	array<H3GenerationReference>	否	H3 全能参考：无视频最多 9 图，有 1 个视频时最多 8 图；最多 3 段音频，总计不超过 12 项且至少一项视觉素材。同类型顺序决定编号；素材 ID 不可重复，且须属于当前调用身份和模型。；最多项数：12

字段	类型	必填	说明
reference_image_url	string	否	可选：与 reference_image_asset_id 并列的外部来源，二选一。只接受公网 HTTPS（443 端口），另接受 图片数据；平台在接纳前经受控下载器摄入为当前调用身份和模型的一小时临时素材，再按资源 ID 原链路处理，等同先调用 /v1/assets/input。同槽位同时传 ID 与 URL 返回 400 ambiguous_input_source；同请求重复 URL 只下载一次；来源被拒为 400 invalid_source_url，下载失败 502 source_fetch_failed，超时 504 source_fetch_timeout；部署关闭本能力时返回 400 input_url_not_enabled。
reference_image_urls	array<string>	否	有序 URL 数组，按位置改写为 reference_image_asset_ids。可选：与 reference_image_asset_ids 并列的外部来源，二选一。只接受公网 HTTPS（443 端口），另接受 图片数据；平台在接纳前经受控下载器摄入为当前调用身份和模型的一小时临时素材，再按资源 ID 原链路处理，等同先调用 /v1/assets/input。同槽位同时传 ID 与 URL 返回 400 ambiguous_input_source；同请求重复 URL 只下载一次；来源被拒为 400 invalid_source_url，下载失败 502 source_fetch_failed，超时 504 source_fetch_timeout；部署关闭本能力时返回 400 input_url_not_enabled。
resolutionTier	string	否	H3 支持 544p、768p；其他服务使用各自公布的清晰度。

字段	类型	必填	说明
seconds	integer	否	请求生成时长，单位秒。H3 为 5 ~ 15 的整数；实际产物时长和帧数另由结果确定。
seed	integer	否	可选随机种子，用于可重复输入；同一幂等请求重试时不得改变。；最小值：0；最大值：2147483647
source_video_url	string	否	可选：与 source_video_asset_id 并列的外部来源，二选一。只接受公网 HTTPS（443 端口）；平台在接纳前经受控下载器摄入为当前调用身份和模型的一小时临时素材，再按资源 ID 原链路处理，等同先调用 /v1/assets/input。同槽位同时传 ID 与 URL 返回 400 ambiguous_input_source；同请求重复 URL 只下载一次；来源被拒为 400 invalid_source_url，下载失败 502 source_fetch_failed，超时 504 source_fetch_timeout；部署关闭本能力时返回 400 input_url_not_enabled。
speedMultiplier	number	否	Wan2.2 Animate 处理速度。720p 支持 1 或 1.5；480p 使用 1。1.5 倍按对应价格系数计费，不承诺上传和排队等总耗时同比缩短。其他模型仅在能力合同声明时使用。；可选值：[1,1.5]

字段	类型	必填	说明
storageMode	string	否	平台存储策略，使用顶层字段：默认 personal_30d，将本次主动上传或生成的资源保存到当前用户资源空间，自资源入库起 30 天到期，可在门户“我的资源”改为永久保存，仍占用配额。temporary_1h 自入库起 1 小时到期，仅后台已对所属用户或组织开启临时存储权限时可选；未授权时拒绝，不静默降级。生成请求不改变已引用素材的原到期时间。公网 URL 受控导入的临时输入遵循独立的 1 小时策略，不因本字段进入个人资源空间。；可选值：["personal_30d","temporary_1h"]；默认值："personal_30d"

```
{
  "generationMode": "t2v",
  "model": "minimax-h3",
  "orientation": "landscape",
  "prompt": "雨夜站台，镜头缓慢推进，人物自然递伞后共同离开。",
  "resolutionTier": "768p",
  "seconds": 5,
  "seed": 1234,
  "storageMode": "personal_30d"
}
```

```
{
  "generationMode": "i2v",
  "input_image_asset_id": "input_first",
  "model": "minimax-h3",
  "orientation": "landscape",
  "prompt": "雨夜站台，镜头缓慢推进，人物自然递伞后共同离开。",
  "resolutionTier": "768p",
  "seconds": 5,
  "seed": 1234,
  "storageMode": "personal_30d"
}
```

```
{
  "end_image_asset_id": "input_last",
  "generationMode": "first_last_frame_video",
  "input_image_asset_id": "input_first",
  "model": "minimax-h3",
  "orientation": "landscape",
  "prompt": "雨夜站台，镜头缓慢推进，人物自然递伞后共同离开。",
  "resolutionTier": "768p",
  "seconds": 5,
  "seed": 1234,
  "storageMode": "personal_30d"
}
```

```
{
  "generationMode": "universal_reference_video",
  "model": "minimax-h3",
  "orientation": "landscape",
  "prompt": "雨夜站台，镜头缓慢推进，人物自然递伞后共同离开。",
  "referenceInputs": [
    {
      "assetId": "input_storyboard",
      "mediaType": "image",
      "role": "reference_image"
    }
  ],
  "resolutionTier": "768p",
  "seconds": 5,
  "seed": 1234,
  "storageMode": "personal_30d"
}
```

允许扩展属性；结构：object

MediaJob

类型：object

字段	类型	必填	说明
assets	array<MediaResultAsset>	否	兼容结果集合；新客户端优先读取 outputs。
clientRequestId	string	否	可选客户端关联标识；不等同于幂等键。
completedAt	string	否	任务完成时间，采用带时区的日期时间。；格式：date-time
createdAt	string	否	创建时间，采用带时区的日期时间。；格式：date-time

字段	类型	必填	说明
details	object	否	可安全公开的审核原因和用户操作提示；不包含原文、候选或内部错误。
details.category	string	否	审核类别。
details.guidanceCode	string	否	受控的用户操作指导代码。
details.reasonCode	string	否	受控原因代码。
details.stage	string	否	审核发生于 input、output 或 generation 阶段。
effectiveMediaSpec	EffectiveMediaSpec	否	类型定义：EffectiveMediaSpec
effectiveSpec	EffectiveMediaSpec	否	类型定义：EffectiveMediaSpec
elapsedSeconds	number	否	任务墙钟总耗时，单位秒；不是计费生成时长。；最小值：0
errorCode	string	否	执行失败时返回的机器错误码。提示词优化技术失败为 prompt_optimization_failed，准备超时为 prompt_optimization_timeout；安全拒绝、澄清、候选拒绝和审核不可用分别保留 content_safety_blocked、content_safety_requires_clarification、prompt_optimization_candidate_rejected、content_safety_unavailable。
errorMessage	string	否	可安全展示的失败说明。
failureStage	string	否	自动提示词准备失败为 prompt_optimization；表示失败发生的位置。
id	string	否	jobId 的兼容别名。
jobId	string	是	受理后持久保存的任务 ID；用 GET /v1/jobs/{id}?model=... 查询原任务。
jobUrl	string	是	带鉴权的绝对轮询地址；携带原 API Key 原样请求，不用 publicationId 自行拼接。；格式：uri

字段	类型	必填	说明
model	string	否	公开模型别名。
object	string	否	通常为 media.generation.job；图像续写任务为 image_continuation.task。
outputs	array<MediaResultAsset>	否	成功后的权威结果集合；优先于兼容结果字段使用。
phase	string	否	自动提示词准备失败的原因类别；安全审核错误与技术异常分开。；可选值: ["prompt_optimization","content_safety"]
promptOptimization	object	否	仅自动优化 Job 返回；缺失表示该 Job 没有自动准备。准备未成功时不进入 H3 Encode。
promptOptimization.attempt	integer	否	当前常规优化 attempt 序号，从 1 开始，最多为 2。；最小值：1；最大值：2
promptOptimization.enabled	boolean	否	该 Job 是否启用持久自动提示词准备。；固定值：true
promptOptimization.maxModelCalls	integer	否	包含每个 attempt 最多一次受控压缩的计费模型调用上界，自动准备为四次。；固定值：4
promptOptimization.maxOptimizationAttempts	integer	否	自动准备最多两个常规优化 attempt；未知结果恢复同一 attempt。；固定值：2
promptOptimization.status	string	否	对象当前状态，必须按所属操作的状态集合判断。；可选值: ["pending","processing","resolving","succeeded","failed"]
queuedSeconds	number	否	排队耗时，单位秒。；最小值：0
refreshAfterSeconds	integer	否	建议的下次刷新间隔，单位秒。；最小值：0
requestId	string	是	服务端生成的请求追踪标识。
requestedSpec	RequestedMediaSpec	否	类型定义: RequestedMediaSpec

字段	类型	必填	说明
retryAfterSeconds	integer	否	建议重试或轮询前等待的秒数。；最小值：0
retryable	boolean	否	失败任务的重试提示；false 表示客户端不得自动重新创建生成任务。
runningSeconds	number	否	执行耗时，单位秒。；最小值：0
serviceKind	string	否	所选服务的公开能力类别。
stage	string	否	任务当前阶段；HTTP 200 不代表生成成功。
stageMessage	string	否	可读的当前阶段说明，不用于驱动状态分支。
startedAt	string	否	任务开始执行时间。；格式：date-time
status	string	是	H3 不提供用户取消接口；兼容状态中的取消值不表示开放了取消能力。；可选值：["queued","running","succeeded","failed","cancel_requested","canceled"]
statusCode	integer	否	关联的 HTTP 状态码。


```
{
  "id": "job_example",
  "jobId": "job_example",
  "jobUrl": "https://api.example.test/v1/jobs/job_example?model=minimax-h3",
  "model": "minimax-h3",
  "requestId": "req_example",
  "serviceKind": "video_generation",
  "stage": "queued",
  "status": "queued"
}
```


允许扩展属性；结构：object

MediaResultAsset

类型：`object`

字段	类型	必填	说明
assetId	string	是	用于 GET/HEAD /v1/assets/{id}/content?model=... 的结果素材 ID；不是任务 ID。
byteSize	integer	否	文件大小，单位字节。；最小值：0
contentUrl	string	是	带鉴权的绝对下载地址，需原 API Key；不是公开或永久分享链接。；格式：uri
expiresAt	string	否	资源当前的自动到期时间；到期后停止用户访问，物理清理由后台异步完成。永久保存时可省略；历史资源缺失时不表示默认 30 天。查询和下载不续期。；格式：date-time
ingestedAt	string	否	结果资源完成入库的时间；保留期限从此计时，不从提交或查询任务时计时。历史资源可能未记录。；格式：date-time
kind	string	否	结果素材的媒体类别。
mimeType	string	否	检测到的 MIME 类型。
sha256	string	否	返回时可用于校验内容完整性的摘要。
storageMode	string	否	该资源实际采用的存储模式；历史资源可能未记录，不应据此补造 30 天期限。永久保存及实际到期时间以资源当前保留策略和 expiresAt 为准。；可选值：["personal_30d","temporary_1h"]

允许扩展属性；结构：object

MiniMaxV2Error

类型：`object`

字段	类型	必填	说明
error	object	是	失败原因对象；任务失败与 HTTP 请求失败分别读取对应对象。
error.code	string	是	稳定机器错误码，供程序分类处理。
error.http_code	string	是	HTTP 状态码的字符串形式。
error.message	string	是	可读说明或消息对象，按所属响应结构解析。
error.type	string	是	当前对象、消息片段或工具的类型。
request_id	string	是	本次请求追踪标识，联系支持时提供。
type	string	是	当前对象、消息片段或工具的类型。；可选值：["error"]

MiniMaxV2Request

类型：`object`

字段	类型	必填	说明
aigc_watermark	boolean	否	兼容字段，仅支持 false。；固定值：false
content	array<object / object / object / object>	是	恰好一项 text。文生无素材；图生 first_frame；首尾帧 first_frame + last_frame；全能参考角色不得与帧角色混用，至少一个视觉素材。最多9图（有视频8图）、1视频、3音频，总素材12项。；最少项数：1；最多项数：13
duration	integer	是	媒体时长，单位秒。；最小值：5；最大值：15
model	string	是	MiniMax V2 固定兼容模型名 MiniMax-H3，区别于原生 V1 模型别名；大小写须一致。；可选值：["MiniMax-H3"]

字段	类型	必填	说明
optimizePrompt	boolean	否	默认先优化再生成，优化单独计费。false 跳过优化；失败不回退原文生成。；默认值：true
prompt_optimizer	boolean	否	optimizePrompt 的别名；同时出现时必须一致。
ratio	string	是	六档固定画幅：21:9、16:9、4:3、1:1、3:4、9:16；新请求必须显式提供。历史任务可能保留旧值。；可选值：["21:9","16:9","4:3","1:1","3:4","9:16"]
resolution	string	是	V2 清晰度，仅接受大小写精确的 768P。；可选值：["768P"]
storageMode	string	否	平台存储策略，使用顶层字段：默认 personal_30d，将本次主动上传或生成的资源保存到当前用户资源空间，自资源入库起 30 天到期，可在门户“我的资源”改为永久保存，仍占用配额。temporary_1h 自入库起 1 小时到期，仅后台已对所属用户或组织开启临时存储权限时可选；未授权时拒绝，不静默降级。生成请求不改变已引用素材的原到期时间。公网 URL 受控导入的临时输入遵循独立的 1 小时策略，不因本字段进入个人资源空间。；可选值：["personal_30d","temporary_1h"]；默认值："personal_30d"

MiniMaxV2Task

类型：`object`

字段	类型	必填	说明
task	object	是	V2 异步任务；必须检查 status，HTTP 200 不代表生成成功。
task.content	object	否	消息内容；支持字符串或模型允许的多模态片段。

字段	类型	必填	说明
task.content.url	string	是	成功时返回的完整票据下载 URL；无需额外 Bearer，不向下载地址转发 API Key，票据最长成功后 7 天有效；资源本身到期或删除后停止下载，包括 temporary_1h 的 1 小时期限。；格式：uri
task.created_at	integer	否	Unix 秒
task.duration	integer	是	媒体时长，单位秒。
task.error	object	否	失败原因对象；任务失败与 HTTP 请求失败分别读取对应对象。
task.error.code	string	是	稳定机器错误码，供程序分类处理。
task.error.message	string	是	可读说明或消息对象，按所属响应结构解析。
task.height	integer	是	高度，单位像素。
task.id	string	是	当前对象的唯一标识。
task.modality	string	否	结果模态，本接口为 video。；可选值：["video"]
task.model	string	是	MiniMax V2 固定兼容模型名 MiniMax-H3，区别于原生 V1 模型别名；大小写须一致。；可选值：["MiniMax-H3"]
task.ratio	string	是	六档固定画幅：21:9、16:9、4:3、1:1、3:4、9:16；新请求必须显式提供。历史任务可能保留旧值。
task.resolution	string	是	V2 清晰度，仅接受大小写精确的 768P。；可选值：["768P"]
task.status	string	是	对象当前状态，必须按所属操作的状态集合判断。；可选值：["queued","running","succeeded","failed","cancelled"]
task.task_type	string	否	任务类别，本接口为 generation。；可选值：["generation"]
task.updated_at	integer	否	任务更新时间，Unix 秒。

字段	类型	必填	说明
task.usage	object	否	本次调用用量；不等同于最终价格。
task.usage.input_image_count	integer	否	图片输入数量，不是计费金额。
task.usage.input_seconds	number	否	参考视频时长
task.usage.output_seconds	number	否	探测的实际产物时长，不是请求时长或金额
task.usage.total_seconds	number	否	参考视频与实际产物时长之和；不包含音频参考时长，不等于平台账单。
task.width	integer	是	宽度，单位像素。

OCRRequest

类型：`object`

字段	类型	必填	说明
document_asset_id	string	否	仅用于明确支持该字段的 OCR 服务；不同于文档解析的素材字段。
image_asset_id	string	否	计量兼容的输入素材标识。
input_image_asset_id	string	否	图片 OCR 推荐使用 /assets/input 上传所得图片 ID；是否必填由具体模型决定。

字段	类型	必填	说明
input_image_url	string	否	可选：与 input_image_asset_id / image_asset_id / document_asset_id 并列的外部来源，二选一。只接受公网 HTTPS（443 端口），另接受 图片数据；平台在接纳前经受控下载器摄入为当前调用身份和模型的一小时临时素材，再按资源 ID 原链路处理，等同先调用 /v1/assets/input。同槽位同时传 ID 与 URL 返回 400 ambiguous_input_source；同请求重复 URL 只下载一次；来源被拒为 400 invalid_source_url，下载失败 502 source_fetch_failed，超时 504 source_fetch_timeout；部署关闭本能力时返回 400 input_url_not_enabled。
model	string	是	OCR 服务公开别名；实际图片格式及任务字段由模型能力决定。
pages	integer	否	计量页数；省略或非正数时默认为 1，不表示允许任意指定执行页数。
task	string	否	模型支持的识别任务，例如 ocr。

OCRResponse

返回所选 OCR 模型的原始 JSON，不采用统一 data 包装。text、blocks 仅代表常见输出，不保证每个模型都有；以服务输出合同为准。

类型：`object`

字段	类型	必填	说明
blocks	array<object>	否	模型识别出的内容块集合。
blocks[].bbox	array<number>	否	识别区域坐标；坐标顺序和单位由模型输出合同确定。

字段	类型	必填	说明
blocks[].confidence	number	否	模型置信度；不是跨模型可直接比较的概率。
blocks[].text	string	否	文本内容。
pages	array<object>	否	输入页数或计量页数；须与文件事实一致。
pages[].blocks	array<object>	否	模型识别出的内容块集合。
pages[].blocks[].bbox	array<number>	否	识别区域坐标；坐标顺序和单位由模型输出合同确定。
pages[].blocks[].text	string	否	文本内容。
pages[].page	integer	否	当前页码，基准见该操作分页参数。
pages[].text	string	否	文本内容。
text	string	否	文本内容。

允许扩展属性；结构：object

OpenAPIDocument

类型：`object`

字段	类型	必填	说明
components	object	否	OpenAPI 可复用类型、参数和认证定义。
info	object	是	OpenAPI 文档标题、版本和说明。
info.description	string	否	面向调用方的说明。
info.title	string	否	文档或内容标题。
info.version	string	否	当前对象的版本。
openapi	string	是	OpenAPI 规范版本，不是 SDK 包版本。
paths	object	是	公开 HTTP 路径及操作定义。
servers	array<object>	否	本协议允许使用的基础地址。

字段	类型	必填	说明
servers[].url	string	否	当前资源地址；携带凭据前验证可信来源。

OrganizationCompute

组织专属算力：每个专属集群一项（集群共享策略为按组织绑定且仅绑定本组织）；没有专属集群时 data 为空数组。观测超过新鲜期时 freshness=stale、health=unknown，指标为 null，不显示为 0 或健康。不返回内部地址、节点名、凭据或调度参数。

类型：`object`

字段	类型	必填	说明
data	array<object>	是	专属集群列表，专属在前、调试在后，按名称排序。
data[].capacity	object	是	剩余容量；观测陈旧时各项为 null。
data[].capacity.busyDevices	integer / null	是	已占满卡数。；格式：int64
data[].capacity.freeMemoryBytes	integer / null	是	可调度卡的剩余可分配显存（字节）。；格式：int64
data[].capacity.idleDevices	integer / null	是	空闲可调度卡数。；格式：int64
data[].capacity.partialDevices	integer / null	是	部分占用卡数。；格式：int64
data[].capacity.unavailableDevices	integer / null	是	不可调度卡数。；格式：int64
data[].debug	boolean	是	是否为调试观测投影的共用集群；其数据为整个共用集群的真实观测，非专属算力。
data[].devices	object	是	卡型与卡数。
data[].devices.healthy	integer / null	是	健康卡数；观测陈旧时为 null。；格式：int64
data[].devices.models	array<object>	是	按卡型汇总，数量多者在前。
data[].devices.models[].count	integer	是	卡数。；格式：int64
data[].devices.models[].deviceType	string	是	设备类型，例如 npu、gpu。

字段	类型	必填	说明
data[].devices.models[].model	string	是	型号。
data[].devices.models[].vendor	string	是	厂商。
data[].devices.total	integer	是	集群卡数（最近一次观测的清单）。；格式：int64
data[].freshness	string	是	观测新鲜度；stale 时健康为 unknown、指标为 null。；可选值：["fresh","stale"]
data[].health	string	是	健康：healthy 全部正常；degraded 部分卡异常或治理中；unavailable 离线或停用；unknown 观测陈旧。；可选值：["healthy","degraded","unavailable","unknown"]
data[].id	string	是	集群不透明标识，同一组织内稳定；不是平台内部集群 ID。
data[].name	string	是	集群展示名称。
data[].nodes	array<object>	是	节点表，按序号排列。
data[].nodes[].computePercent	number / null	是	节点平均算力利用率（%）；无当前遥测时为 null。；最小值：0；最大值：100
data[].nodes[].deviceCount	integer	是	节点卡数。；格式：int64
data[].nodes[].health	string	是	健康：healthy 全部正常；degraded 部分卡异常或治理中；unavailable 离线或停用；unknown 观测陈旧。；可选值：["healthy","degraded","unavailable","unknown"]
data[].nodes[].healthyDevices	integer / null	是	健康卡数；观测陈旧时为 null。；格式：int64
data[].nodes[].id	string	是	节点不透明标识，同一组织内稳定。
data[].nodes[].idleDevices	integer / null	是	空闲卡数；观测陈旧时为 null。；格式：int64
data[].nodes[].label	string	是	展示标签，例如“节点 1”。

字段	类型	必填	说明
data[].nodes[].memoryPercent	number / null	是	节点显存占用率（%）；无当前遥测时为 null。；最小值：0；最大值：100
data[].nodes[].observedAt	string / null	是	节点最近心跳时间；未知时为 null。；格式：date-time
data[].observedAt	string / null	是	集群最近心跳时间；未知时为 null。；格式：date-time
data[].regionName	string	是	区域名称；未设置时为空字符串。
data[].services	array / null	是	本组织在该集群上的服务副本；路由视图不可读时为 null。
data[].services[].desired	integer	是	期望副本数（应提供服务的后端）。
data[].services[].ready	integer / null	是	就绪副本数（活跃且健康观测新鲜）；任一副本健康观测陈旧时为 null。
data[].services[].service	string	是	公开模型别名。
data[].services[].serviceKind	string	是	服务能力类别。
data[].servicesReason	string / null	是	services 为 null 的原因。；可选值：["routing_unavailable",null]
data[].utilization	object	是	利用率与显存占用；只统计有当前遥测的卡。
data[].utilization.computePercent	number / null	是	平均算力利用率（%）；无当前遥测时为 null。；最小值：0；最大值：100
data[].utilization.memoryPercent	number / null	是	显存占用率（%）；无当前遥测时为 null。；最小值：0；最大值：100
data[].utilization.memoryTotalBytes	integer	是	集群显存总量（字节）。；格式：int64
data[].utilization.memoryUsedBytes	integer / null	是	显存占用字节数；无当前遥测时为 null。；格式：int64
data[].utilization.telemetryDevices	integer	是	提供当前遥测的卡数；小于 devices.total 时利用率只代表这部分卡。；格式：int64

字段	类型	必填	说明
debug	boolean	是	是否含调试观测数据（开发环境把共用集群按专属投影）；为 true 时门户须标注“调试数据：共用集群，非专属算力”。
generatedAt	string	是	投影生成时间。；格式：date-time
object	string	是	对象类型。；固定值："organization.compute"

OrganizationContext

当前 Account API Key 所属组织与身份；仅组织当前主账号可读。

类型：`object`

字段	类型	必填	说明
account	object	是	Key 所属账号。
account.displayName	string / null	是	账号显示名称；目录暂不可用时为 null。
account.id	string	是	账号标识。
account.oidcSubject	string / null	是	账号在平台 OIDC 中的主体标识，与 id_token 的 sub 相同；门户据此判断登录人是否为当前主账号。未知时为 null。
apiKey	object	是	当前使用的 Key。
apiKey.id	string	是	Key 标识。
apiKey.scopeType	string	是	Key 授权范围类型。
isPrimary	boolean	是	是否为组织当前主账号；非主账号请求直接返回 403。；固定值：true
object	string	是	对象类型。；固定值："organization.context"
organization	object	是	Key 所属客户组织。
organization.id	string	是	组织标识。
organization.name	string	是	组织名称。

字段	类型	必填	说明
organization.status	string	是	组织状态；可读取时为 active。
workspaceId	string	是	Key 所属个人工作空间，也是计费 owner。

OrganizationPortalUser

中心保存的门户用户映射；门户用户不是平台账号。

类型：`object`

字段	类型	必填	说明
displayName	string	是	门户内的账号显示名。
object	string	是	对象类型。；固定值："organization.portal_user"
portalUserId	string	是	门户用户标识，与调用时 user 字段一致。
registeredAt	string	是	门户用户注册时间。；格式：date-time
status	string	是	门户用户状态。；可选值：["active","frozen"]
updatedAt	string	是	中心最近一次写入时间。；格式：date-time

OrganizationPortalUserRequest

门户用户映射；只接受以下字段，手机号等个人信息不得上送。

类型：`object`

字段	类型	必填	说明
displayName	string	是	门户内的账号显示名。；最长长度：128
registeredAt	string	是	门户用户注册时间。；格式：date-time
status	string	是	门户用户状态。；可选值：["active","frozen"]

OrganizationQuota

主账号在平台侧的准入上限、来源与当前占用。上限按“用户覆盖 → 组织覆盖 → 平台默认”取首个显式值，0 表示禁止；占用读不到时为 null 并给出原因，不显示为 0。同步调用不设账号并发上限，只限每秒新请求数。

类型：`object`

字段	类型	必填	说明
mediaQueue	object	是	异步媒体任务排队份额（单位）：已提交、尚未开始运行的任务占用。
mediaQueue.inUse	integer / null	是	当前占用；读不到时为 null。；格式：int64
mediaQueue.inUseReason	string / null	是	inUse 为 null 的原因；可读时为 null。；可选值：["usage_unavailable",null]
mediaQueue.jobs	integer / null	是	占用该份额的任务数；读不到时为 null。；格式：int64
mediaQueue.limit	integer	是	生效上限；0 表示禁止新请求或新任务。；最小值：0；格式：int64
mediaQueue.source	string	是	上限来源：account 为该账号覆盖，organization 为组织覆盖，platform 为平台默认。；可选值：["account","organization","platform"]
mediaQueue.unit	string	是	上限与占用的单位。；固定值："units"
mediaRunning	object	是	异步媒体任务运行份额（单位）：正在运行的任务占用。
mediaRunning.inUse	integer / null	是	当前占用；读不到时为 null。；格式：int64
mediaRunning.inUseReason	string / null	是	inUse 为 null 的原因；可读时为 null。；可选值：["usage_unavailable",null]
mediaRunning.jobs	integer / null	是	占用该份额的任务数；读不到时为 null。；格式：int64
mediaRunning.limit	integer	是	生效上限；0 表示禁止新请求或新任务。；最小值：0；格式：int64

字段	类型	必填	说明
mediaRunning.source	string	是	上限来源：account 为该账号覆盖，organization 为组织覆盖，platform 为平台默认。；可选值：["account","organization","platform"]
mediaRunning.unit	string	是	上限与占用的单位。；固定值："units"
object	string	是	对象类型。；固定值："organization.quota"
observedAt	string	是	占用观测时间。；格式：date-time
policyRevision	integer	是	生效规则版本；任一层规则变化都会递增。；格式：int64
requestRate	object	是	同步模型调用每秒新请求数：全部 Key、页面与受托应用的同账号调用共享过去 1000ms 滑动窗口；目录、轮询、取消、上传与异步任务提交不计入。
requestRate.inUse	integer / null	是	当前占用；读不到时为 null。；格式：int64
requestRate.inUseReason	string / null	是	inUse 为 null 的原因；可读时为 null。；可选值：["usage_unavailable",null]
requestRate.limit	integer	是	生效上限；0 表示禁止新请求或新任务。；最小值：0；格式：int64
requestRate.source	string	是	上限来源：account 为该账号覆盖，organization 为组织覆盖，platform 为平台默认。；可选值：["account","organization","platform"]
requestRate.unit	string	是	上限与占用的单位。；固定值："requests_per_second"
requestRate.windowMs	integer	是	滑动窗口长度（毫秒）。；固定值：1000

字段	类型	必填	说明
services	array<object>	是	本组织可调用的异步媒体服务一次任务占用的份额，与平台准入扣减同源；门户据此把份额上限换算为可同时排队与运行的任务数。无法确定份额的服务省略。
services[].queueUnits	integer	是	一次任务占用的排队份额（单位）。；最小值：1；格式：int64
services[].runUnits	integer	是	一次任务占用的运行份额（单位）。；最小值：1；格式：int64
services[].service	string	是	公开模型别名。
services[].serviceKind	string	是	服务类型，例如 image_generation、video_generation。

OrganizationServiceStatusList

组织可调用服务的状态分档；不包含内部地址或共享集群负载数字。

类型：`object`

字段	类型	必填	说明
data	array<object>	是	每个服务一项；未传 services 时按服务名排序，传入时按传入顺序。
data[].availability	object	是	近 24 小时（24 个小时槽）与近 7 天（7 个天槽）的可用性。
data[].availability.last24h	array<object>	是	按时间升序的小时槽，最后一个为当前小时。
data[].availability.last24h[].availability	number / null	是	非平台故障调用占比（1 - 平台侧失败 / 调用），四位小数；无调用或未观测时为 null。；最小值：0；最大值：1
data[].availability.last24h[].start	string	是	时间槽开始时间（小时槽按整点，天槽按 Asia/Shanghai 零点）。；格式：date-time

字段	类型	必填	说明
data[].availability.last24h[].status	string	是	该时间槽的分档；no_data 为无调用，unknown 为观测不覆盖。；可选值：["normal","busy","degraded","unavailable","maintenance","unknown","no_data"]
data[].availability.last7d	array<object>	是	按时间升序的天槽，最后一个为今天。
data[].availability.last7d[].availability	number / null	是	非平台故障调用占比（1 - 平台侧失败 / 调用），四位小数；无调用或未观测时为 null。；最小值：0；最大值：1
data[].availability.last7d[].start	string	是	时间槽开始时间（小时槽按整点，天槽按 Asia/Shanghai 零点）。；格式：date-time
data[].availability.last7d[].status	string	是	该时间槽的分档；no_data 为无调用，unknown 为观测不覆盖。；可选值：["normal","busy","degraded","unavailable","maintenance","unknown","no_data"]
data[].observedAt	string / null	是	判定所依据观测中较旧的一个的时间；无观测时为 null。
data[].reasons	array<string>	是	判定依据代码，例如 no_healthy_backend、partial_backends_unhealthy、recent_failure_rate_high、capacity_rejection_rate_high、backend_health_stale、call_observation_stale、service_not_listed。
data[].routePool	string	是	dedicated 为组织专属集群，shared 为共享集群；不在可调用集合中的服务为空串。
data[].service	string	是	公开模型别名。
data[].serviceKind	string	是	服务能力类别；不在本组织可调用集合中的服务为空串。

字段	类型	必填	说明
data[].status	string	是	当前分档：unavailable 无可路由健康后端；degraded 部分后端不健康或近 5 分钟失败率超阈值；busy 容量拒绝或排队超阈值；maintenance 下架、维护或不再向本组织开放；unknown 观测陈旧。；可选值：["normal","busy","degraded","unavailable","maintenance","unknown"]
data[].summary	string	是	判定依据中文摘要。
generatedAt	string	是	生成时间。；格式：date-time
object	string	是	对象类型。；固定值："organization.services.status"

OrganizationUsageMetrics

主账号个人工作空间已结束调用按服务与时间桶的聚合；空桶省略。

类型：`object`

字段	类型	必填	说明
data	array<object>	是	每个服务一项，按服务名排序。
data[].buckets	array<object>	是	按时间升序的非空时间桶。
data[].buckets[].calls	integer	是	已结束的调用 attempt 数（成功与各类失败合计）。；最小值：0；格式：int64
data[].buckets[].failures	object	是	失败分类计数。
data[].buckets[].failures.capacity	integer	是	平台容量不足或排队拒绝。；最小值：0；格式：int64
data[].buckets[].failures.insufficient_balance	integer	是	余额不足拒绝。；最小值：0；格式：int64
data[].buckets[].failures.other	integer	是	其他失败，含参数错误与客户端取消。；最小值：0；格式：int64

字段	类型	必填	说明
data[].buckets[].failures.rate_limited	integer	是	Key 或账号的速率、配额或并发限制。；最小值：0；格式：int64
data[].buckets[].failures.upstream	integer	是	上游运行时或后端失败、超时。；最小值：0；格式：int64
data[].buckets[].firstTokenLatencyMs	object	是	时延分位（毫秒）；无事实时为 null。
data[].buckets[].firstTokenLatencyMs.p50	number / null	是	P50 毫秒。
data[].buckets[].firstTokenLatencyMs.p95	number / null	是	P95 毫秒。
data[].buckets[].outputTokensPerSecond	number / null	是	成功调用的输出 token 合计 / 其总时长合计（tokens/s）；非 token 计量服务为 null。
data[].buckets[].queueWaitMs	object / null	否	媒体任务受理到开始执行的等待分位（毫秒）；非媒体服务或无观测时分位为 null，旧版治理服务不返回时省略。
data[].buckets[].queueWaitMs.p50	number / null	是	P50 毫秒。
data[].buckets[].queueWaitMs.p95	number / null	是	P95 毫秒。
data[].buckets[].start	string	是	时间桶开始时间（Asia/Shanghai 对齐）。；格式：date-time
data[].buckets[].succeeded	integer	是	成功调用数。；最小值：0；格式：int64
data[].buckets[].successRate	number / null	是	成功率，四位小数；calls 为 0 时为 null。；最小值：0；最大值：1
data[].buckets[].totalLatencyMs	object	是	时延分位（毫秒）；无事实时为 null。
data[].buckets[].totalLatencyMs.p50	number / null	是	P50 毫秒。
data[].buckets[].totalLatencyMs.p95	number / null	是	P95 毫秒。
data[].service	string	是	公开模型别名。

字段	类型	必填	说明
data[].serviceKind	string	是	服务能力类别。
from	string	是	窗口下界（含）。；格式: date-time
granularity	string	是	时间桶粒度。；可选值: ["hour","day"]
object	string	是	对象类型。；固定值: "organization.usage_metrics"
observedAt	string	是	聚合读取时间。；格式: date-time
timezone	string	是	时间桶对齐时区。；固定值: "Asia/Shanghai"
to	string	是	窗口上界（不含）。；格式: date-time

PromptOptimizationMediaSpec

仅解析列出的规格维度；哪些必填由所选生成模式决定。

类型: `object`

字段	类型	必填	说明
aspectRatio	string	否	受控输出长宽比；H3 由服务能力合同返回，例如 21:9、16:9、4:3、1:1、3:4、9:16。
imageCount	integer	否	请求生成图片数量；以模型允许范围为准。
orientation	string	否	画幅方向，合法值按模型规格选择。
outputSizeCode	string	否	模型公布的输出尺寸代码；不要自行拼接尺寸。
qualityTier	string	否	模型公布的质量档位。
resolutionTier	string	否	所选模型支持的清晰度档位。
seconds	integer	否	请求时长，单位秒；仅适用于声明支持该字段的模型。

PromptOptimizationRequest

类型：`object`

字段	类型	必填	说明
audioReferencePolicy	string	否	H3 存在音频参考时必填。优化器只理解图片/视频；音频仅按明确文字意图保留用途，不分析其内容。；可选值：["preserve_without_understanding"]
clientRequestId	string	是	客户端业务请求标识；优化操作中须与幂等键一致。；最短长度：8；最长长度：128；匹配规则：^[A-Za-z0-9._:-]+$
dialogueLanguage	string	否	可选台词语种。auto（默认）表示由提示词语言自动解析；其余值为官方稳定支持的台词语言。分角色语种请在提示词中写明。仅 H3 typed-media/v1 支持。；可选值：["auto","arabic","chinese","english","french","german","italian","japanese","korean","portuguese","russian","spanish"]
generationMode	string	是	允许的模式来自所选服务的 promptOptimization.profiles。
mediaSpec	PromptOptimizationMediaSpec	是	类型定义：PromptOptimizationMediaSpec
model	string	是	GET /v1/services 返回的公开模型别名。
originalPrompt	string	是	Qwen Image 2.1 支持最多 10000 个 Unicode 字符；H3 typed-media/v1 仍最多 7000，其他优化配置按所选服务公布的字符上限。；最短长度：1；最长长度：10000

字段	类型	必填	说明
referenceInputs	array<object>	是	使用当前身份上传的素材 ID。H3 角色为 first_frame/last_frame，或 reference_image（0～9）、reference_video（0～1）、reference_audio（0～3）；全能参考至少一项视觉素材，总计不超过 12。同类型顺序决定编号。Qwen Image 2.1 文生图为空数组，编辑为有序 reference_image_1..reference_image_10（1～10 张）；其他配置使用各自角色合同。；最多项数：12
referenceInputs[].assetId	string	是	素材唯一标识，不是任务 ID。
referenceInputs[].role	string	是	消息或素材的角色，枚举含义依所属对象确定。
style	string	否	H3 typed-media/v1 的可选创作风格偏好。；最短长度：1；最长长度：200

PromptOptimizationResponse

类型：`object`

字段	类型	必填	说明
optimization	object	是	本次提示词优化结果。
optimization.createdAt	string	是	创建时间，采用带时区的日期时间。；格式：date-time
optimization.expiresAt	string	是	到期时间；到期后不可假定该资源仍有效。；格式：date-time
optimization.generationMode	string	是	生成模式，合法值及条件字段由模型合同声明。
optimization.noticeCodes	array<string>	否	可供客户端展示对应提示的机器通知码集合。
optimization.optimizationId	string	是	优化结果标识；生成视频时仍传最终提示词文本。

字段	类型	必填	说明
optimization.optimizedPrompt	string	是	原样可编辑的优化正文；Qwen Image 2.1 最多 10000 个 Unicode 字符，H3 仍最多 7000，其他服务按其优化合同。；最长长度：10000
optimization.originalPrompt	string	是	用户提交的原始提示词。
optimization.outcome	string	否	优化处理结果分类。
optimization.serviceKind	string	是	能力类别，例如 video_generation。
optimization.status	string	是	对象当前状态，必须按所属操作的状态集合判断。；固定值："succeeded"
requestId	string	是	服务端请求追踪标识；报障时提供，勿附真实凭据。

PublicError

类型：`object`

字段	类型	必填	说明
code	string	是	稳定机器错误码，供程序分类处理。
details	object	否	结构化错误详情；只在返回时读取。
error	object	是	标准 SDK 可读取的平台错误正文。
error.code	string	是	稳定机器错误码，供程序分类处理。
error.message	string	是	可读说明或消息对象，按所属响应结构解析。
error.param	string / null	是	发生错误的请求参数路径；非参数错误为空。
error.type	string	是	当前对象、消息片段或工具的类型。
message	string	是	可读说明或消息对象，按所属响应结构解析。

字段	类型	必填	说明
requestId	string	是	服务端生成的请求追踪标识。
retryable	boolean	是	该失败是否允许按当前操作合同重试，不表示可以无条件重新生成。

PublicModel

类型：`object`

字段	类型	必填	说明
id	string	是	当前对象的唯一标识。
mediaSpecContract	object	否	当前模型的输入、输出及规格约束。
mediaSpecContract.parameters	array<object>	否	工具或协议的参数定义。
mediaSpecContract.parameters[].description	string	否	面向调用方的说明。
mediaSpecContract.parameters[].name	string	否	当前对象或工具的名称。
mediaSpecContract.parameters[].required	boolean	否	必填属性名称集合。
mediaSpecContract.parameters[].type	string	否	当前对象、消息片段或工具的类型。
mediaSpecContract.version	string	否	当前对象的版本。
object	string	是	返回对象的类型标记。；固定值："model"
owned_by	string	是	模型提供方或所属主体的公开标识。

PublicModelList

类型：`object`

字段	类型	必填	说明
data	array<PublicModel>	是	本操作的数据集合；不同接口不保证共用同一种信封。
object	string	是	返回对象的类型标记。；固定值："list"

PublicService

类型：`object`

字段	类型	必填	说明
apiUsageContract	object	否	所选模型当前公开的操作、字段和示例。
automaticPromptOptimization	object	否	仅公开 Account API Key H3 的自动准备能力；可用性表示部署激活及本地编排配置，不代表模型实时健康。
automaticPromptOptimization.available	boolean	是	仅公开 Account API Key H3：此部署已激活且本地编排配置就绪；不代表模型/后端实时健康。
automaticPromptOptimization.defaultEnabled	boolean	是	此入口省略 optimizePrompt 时是否默认开启自动优化；H3 为 true，部署未激活时返回能力不可用。；固定值：true
automaticPromptOptimization.maxModelCalls	integer	是	包含每个 attempt 最多一次受控压缩的计费模型调用上界，自动准备为四次。；固定值：4
automaticPromptOptimization.maxOptimizationAttempts	integer	是	自动准备最多两个常规优化 attempt；未知结果恢复同一 attempt。；固定值：2
available	boolean	否	当前是否可调用；发现时可用不保证后续准入。
capability	object	否	服务能力投影，例如上下文长度与输入模态。
endpoints	object	否	服务公开支持的接口地址。
id	string	否	当前对象的唯一标识。
limits	object	否	文本与多模态对话的调用限制；其他服务省略。

字段	类型	必填	说明
limits.defaultMaxOutputTokens	integer	是	请求省略 max_tokens 时平台转发的默认输出上限。；最小值：1；格式：int64
mediaSpecContract	object	否	模型当前的必填字段、允许值、模式与输入组合、媒体大小限制。
modelAlias	string	是	服务发现返回的公开模型别名，用于后续 model 字段。
object	string	否	返回对象的类型标记。
pricing	PublicServicePricing	否	类型定义：PublicServicePricing
promptOptimization	object	否	独立提示词优化的可用性、模式、参考角色/限制和超时建议。
serviceCapabilityDescription	string	否	服务能力说明。
serviceDescription	string	否	服务简介。
serviceDisplayName	string	否	服务的用户可读名称。
serviceKind	string	是	能力类别，例如 video_generation。

PublicServiceList

类型：`object`

字段	类型	必填	说明
data	array<PublicService>	是	本操作的数据集合；不同接口不保证共用同一种信封。
object	string	是	返回对象的类型标记。；固定值："list"

PublicServicePricing

调用方（Key 所属账号及其组织）的生效价格；无法确定时整个对象省略，不返回 0。实际扣费以调用明细为准。

类型：`object`

字段	类型	必填	说明
currency	string	是	计价币种，平台积分。；固定值："credit"
defaultSpec	string	是	价格对应的默认规格编码；无规格区分时为 default。
items	array<object>	是	默认规格的各计量项单价。
items[].meter	string	是	计量项，例如 input_tokens、output_tokens 或媒体资源类型。
items[].unit	string	是	计价单位，例如 1M tokens。
items[].unitPriceMilliPoints	integer	是	每个计价单位的单价（千分之一积分）。；格式：int64
scope	string	是	生效价格规则范围。；可选值：["platform","organization","publication"]
specs	array<object>	是	报价中全部可确定价格的规格，默认规格在首位；无法确定价格的规格省略，不返回 0。单规格服务只有一个默认项。
specs[].code	string	是	规格编码，与调用明细的计价规格一致。
specs[].default	boolean	是	是否为默认规格。
specs[].items	array<object>	是	该规格各计量项单价。
specs[].items[].meter	string	是	计量项，例如 input_tokens、output_tokens 或媒体资源类型。
specs[].items[].unit	string	是	计价单位，例如 1M tokens。
specs[].items[].unitPriceMilliPoints	integer	是	每个计价单位的单价（千分之一积分）。；格式：int64
specs[].label	string	是	规格展示名，取服务 mediaSpecContract 中的规格名；取不到时为规格编码。

RequestedMediaSpec

调用方实际请求的媒体规格；旧请求缺少新字段时不补造。

类型：`object`

字段	类型	必填	说明
aspectRatio	string	否	请求的画幅比例。
generationMode	string	否	请求的生成模式。
orientation	string	否	请求的画幅方向。
outputSizeCode	string	否	请求的固定输出尺寸代码。
resolutionTier	string	否	请求的清晰度档位。

RerankRequest

类型：`object`

字段	类型	必填	说明
documents	array<string>	是	待重排的候选文本集合，顺序用于与结果 index 对齐。；最少项数：1；最多项数：64
model	string	是	重排模型的公开别名。
query	string	是	用于重排候选内容的查询文本。；最短长度：1；最长长度：1024
top_n	integer	否	仅在模型支持时发送；query 与 documents 总计不超过 65536 个 Unicode 字符。；最小值：1

RerankResponse

类型：`object`

字段	类型	必填	说明
id	string	否	当前对象的唯一标识。
model	string	否	服务发现返回的公开模型别名。
results	array<object>	否	本操作返回的结果集合。
results[].document	string / object	否	候选文档或返回的原文信息。

字段	类型	必填	说明
results[].index	integer	否	结果对应的输入或候选位置，从 0 开始。
results[].relevance_score	number	否	相关性分数，用于同次查询排序，不是通用概率。
usage	TokenUsage	否	类型定义：TokenUsage

ResponsesRequest

类型：`object`

字段	类型	必填	说明
background	boolean	否	仅支持 false；调用随当前连接执行。；固定值：false
frequency_penalty	number	否	按出现频率调整 token 的采样惩罚。
include	array<string>	否	请求存在时的可选推理加密字段；当前仅返回可兑现内容。
input	string / array<object>	是	字符串，或 message/function_call/function_call_output/明文 reasoning 历史。图片使用 input_image.image_url，支持所选模型允许的 HTTPS/data URI。
instructions	string	否	放在本轮历史前的系统指令。
max_output_tokens	integer	否	本轮最大输出 token 数，受模型预算合同限制。；最小值：1
metadata	object	否	最多 16 项字符串元数据；键和值分别不超过 64 和 512 字符。
model	string	是	所选模型的公开别名。
parallel_tool_calls	boolean	否	是否允许本轮生成多个函数调用。
presence_penalty	number	否	按是否出现调整 token 的采样惩罚。
prompt_cache_key	string	否	可选的提示缓存分区标识；不保证缓存命中。

字段	类型	必填	说明
reasoning	object	否	模型推理配置。
reasoning.effort	string	否	传给所选模型的推理预算级别。；可选值：["none","minimal","low","medium","high","xhigh","max"]
reasoning.summary	string	否	auto 投影模型已有的明文思考内容；不执行独立摘要生成，不生成加密状态。；可选值：["auto"]
seed	integer	否	可选整数随机种子。
service_tier	string	否	仅支持 auto/default，使用模型默认服务级别。；可选值：["auto","default"]
stop	string / array<string>	否	自定义停止字符串，受模型流式及媒体合同约束。
store	boolean	否	仅支持 false；服务端不保存可续接的 Responses 会话。；固定值：false；默认值：false
stream	boolean	否	是否返回带类型的增量 SSE 事件。
temperature	number	否	采样温度，受模型能力约束。
text	object	否	文本输出格式配置。
text.format	object	否	文本或模型支持的结构化输出格式。
tool_choice	string / object	否	自动、禁止、必需或指定名称的函数调用策略。
tools	array<object>	否	函数工具定义；不支持服务器内置工具。
tools[].description	string	否	工具功能说明。
tools[].name	string	是	函数或结构化输出的名称。
tools[].parameters	object	否	函数参数的 JSON Schema，支持范围由模型合同规定。
tools[].strict	boolean	否	是否严格约束函数参数。
tools[].type	string	是	该对象的协议类型。；固定值："function"

字段	类型	必填	说明
top_p	number	否	累积概率采样阈值。
truncation	string	否	仅支持 disabled；超预算明确拒绝。；固定值："disabled"
user	string	否	调用方用户标识。

ResponsesResponse

类型：`object`

字段	类型	必填	说明
created_at	integer	是	响应创建时的 Unix 秒级时间。
error	object / null	否	执行失败时的错误，成功时为空。
id	string	是	服务端生成的响应标识。
incomplete_details	object / null	否	达到输出预算或内容过滤导致的不完整原因。
model	string	是	所选模型的公开别名。
object	string	是	固定为 response。；固定值："response"
output	array<object>	是	按顺序排列的消息、明文推理和函数调用项。
status	string	是	当前响应的执行状态。；可选值：["in_progress","completed","incomplete","failed"]
store	boolean	否	仅支持 false；服务端不保存可续接的 Responses 会话。；固定值：false
usage	object / null	是	input_tokens/output_tokens/total_tokens 及缓存、reasoning token 明细。

允许扩展属性；结构：object

SynchronousMediaResponse

同步媒体服务返回的原始 JSON，不是异步任务对象。下列可选字段描述常见图片兼容结果，不保证每个服务都有；精确输出以所选服务合同为准。

类型：`object`

字段	类型	必填	说明
created	integer	否	创建时间，通常为 Unix 秒时间戳，以所属对象合同为准。
data	array<object>	否	本操作的数据集合；不同接口不保证共用同一种信封。
data[].b64_json	string	否	服务支持时返回的 Base64 图片内容。
data[].revised_prompt	string	否	服务返回时提供的修订提示词。
data[].url	string	否	当前资源地址；携带凭据前验证可信来源。
usage	object	否	本次调用用量；不等同于最终价格。
usage.total_tokens	integer	否	输入和输出的总 Token 数。

允许扩展属性；结构：object

TokenUsage

类型：`object`

字段	类型	必填	说明
completion_tokens	integer	否	输出 Token 数，不等同于字符数。
prompt_tokens	integer	否	输入 Token 数。
total_tokens	integer	否	输入和输出的总 Token 数。

凭据 bearerAuth


```
{
  "bearerFormat": "Account API Key",
  "scheme": "bearer",
  "type": "http"
}
```


凭据 inputAssetReadToken

平台签发的短期输入读取令牌；原样使用已签发 URL，API Key 不能替代此令牌。


```
{
  "description": "平台签发的短期输入读取令牌；原样使用已签发 URL，API Key 不能替代此令牌。",
  "in": "query",
  "name": "token",
  "type": "apiKey"
}
```


凭据 miniMaxV2Ticket

原样使用查询返回的完整 URL；票据绑定原 API Key 的权限和生命周期。


```
{
  "description": "原样使用查询返回的完整 URL；票据绑定原 API Key 的权限和生命周期。",
  "in": "query",
  "name": "ticket",
  "type": "apiKey"
}
```