TokenHub

产品简介

腾讯云

【联系我们】

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

文档目录

产品简介
产品概述

最近更新时间：2026-04-03 21:02:21

什么是 TokenHub?

TokenHub 是一站式 AI 大模型服务平台，面向企业和开发者提供统一的模型接入能力。平台聚合了腾讯混元、DeepSeek、MiniMax、Kimi、智谱 GLM、通义千问等多家行业领先的大模型，覆盖文本生成、图片生成、视频生成、3D 生成等多种 AI 能力场景。
无论您是希望快速集成 AI 能力的应用开发者，还是需要稳定可靠的模型推理服务的企业用户，TokenHub 都为您提供了一站式解决方案：**统一的 API 接口、灵活的计费模式、完善的监控体系**以及**便捷的在线体验**，帮助您以较低的接入成本获得优质的大模型服务。

核心功能

功能	说明
模型广场	汇聚多家厂商的主力大模型，提供多维度筛选（类型、服务商、体验状态）、模型对比和详细信息查看，帮助您快速选择合适的模型使用。平台提供了部分模型的免费体验包，供用户免费试用平台部分模型。
体验中心	在线体验大模型能力，支持文本对话（深度思考、流式输出）、图片生成（多风格、多尺寸）、视频生成以及 3D 生成，便于您在接入 API 之前充分评估模型效果。
AI 创作	视频特效中心展示了海量创意视频特效模板，支持一键复刻爆款同款和 API 灵活调用，方便根据个人需求选择、体验和调用特效。
在线推理	创建和管理模型的推理服务实例，获取专属 API Endpoint。支持按需选择计费方式——免费体验、按 Token 计费，满足从测试验证到生产部署的全阶段需求。
模型监控	展示模型、服务性能相关指标，如：TTFT (Time To First Token)、TPOT (Time Per Output Token)、RPM (Request Per Minute) 。
用量统计	展示模型、服务性能计费用量指标，如：输入 Token、输出 Token、TPM (Tokens Per Minute) 、插件调用次数。
API Key 管理	集中管理 API 访问密钥，支持精细化的权限控制（全部模型及服务或限定范围），便捷地启停切换和调用统计，保障接口调用的安全性与可控性。
Token Plan	是面向龙虾和编程场景设计的专属订阅套餐，覆盖腾讯混元、MiniMax、GLM、Kimi 等国产主流模型，兼容热门龙虾工具和主流编程工具，内设更多套餐档位可供选择。
Coding	面向开发者的 AI 编程工具套餐，提供 Lite 和 Pro 两档选择，支持 OpenClaw、Cursor

Plan	等常见 AI 工具集成，提升开发效率。

平台支持的模型

- 平台支持的模型清单，请参见 模型列表 。
- 模型计费方式，请参见 计费方式 。

快速开始使用

您可以通过 快速入门 文档了解如何领取体验包、创建服务并调用 API。

模型列表

最近更新时间：2026-04-03 16:27:22

语言模型

模型名称	model（调用参数）	能力支持	上下文窗口（Token）	最大输入（Token）	最大输出（Token）
HY 2.0 Think	hunyuan-2.0-thinking-20251109	• 深度思考 • 联网搜索 • Function Calling	192k	128k	64k
HY 2.0 Instruct	hunyuan-2.0-instruct-20251111	• 联网搜索 • Function Calling	144k	128k	16k
Hunyuan-role	hunyuan-role-latest	角色扮演模型适用 AI 数字分身、AI 角色扮演、AI 情感陪聊等场景	32k	28k	4k
Deepseek-v3.2	deepseek-v3.2	• 深度思考 • 结构化输出 • Function Calling	128k	96k	32k
Deepseek-v3.1	deepseek-v3.1-terminus	• 深度思考 • 结构化输出 • Function Calling	128k	96k	32k
Deepseek-r1-0528	deepseek-r1-0528	• 深度思考 • 结构化输出	128k	96k	16k

		• Function Calling			
Deepseek-v3-0324	deepseek-v3-0324	• Function Calling	128k	128k	16k
GLM-5	glm-5	• 深度思考 • Function Calling • Cache 缓存	200k	200k	128k
kimi-k2.5	kimi-k2.5	• 深度思考 • 结构化输出 • Function Calling • Cache 缓存	256k	224k	16k
MiniMax-M2.5	minimax-m2.5	• 深度思考 • Function Calling • Cache 缓存	200k	200k	128k
MiniMax-M2.7	minimax-m2.7	• 深度思考 • Function Calling • Cache 缓存	200k	200k	128k

视觉模型

图像生成

模型名称	model（调用参数）	模型介绍	任务类型	默认并发数
HY-Image-V3.0	hy-image-v3.0	基于混元大模型，能够去思考图像的布局、构图、笔触，利用世界知识去推理常识性的画面。同时可以解析千字级别的复	• 文生图 • 图生图	1

		杂语义，生成长文本文字、复杂漫画、表情包，还能生成生动有趣的科普插画。		
HY-Image-Lite	hy-image-lite	采用超高压缩编解码器，实现图像生成快速响应与高品质输出。支持电商商品图美化、设计工具素材生成、游戏场景迭代等场景。	文生图	1

视频生成

模型名称	model（调用参数）	模型介绍	任务类型	默认并发数
HY-Video-1.5	hy-video-1.5	支持文本、图像多模态输入生成高清视频，可实现场景切换与多角色交互，简化制作流程、降低成本，应用于企业广告营销与个人创意落地场景。	• 文生视频 • 图生视频	5
YT-Video-2.0	yt-video-2.0	支持生成动态连贯性高、画面过渡自然的视频，适用于对质量要求较高的广告创意、影视片段和产品展示等场景。	图生视频	5
YT-Video-HumanActor	yt-video-humanactor	单张参考照片即可驱动生成动态人像视频，精准还原表情、姿态，支持写实、二次元等多风格切换。	图生视频	5
YT-Video-FX	yt-video-fx	通过上传图片和选择特效模板，生成一段特效视频，将静态图像转化为充满活力、动感、有趣的视频画面。	图生视频	5

3D 生成

模型名称	model（调用参数）	模型介绍	任务类型	默认并发数
HY-3D-3.0	hy-3d-3.0	采用混元生3D 3.0模型，可生成更高精度以及更高质量的3D 模型，支持文生3D、图生3D、多视图生3D、单几何生成（白模），草图生3D、智能拓扑生3D功能。	• 文生3D • 图生3D	3
HY-3D-3.1	hy-3d-3.1	采用混元生3D 3.1模型，可生成更高精度以及更高质量的3D 模型，支持文生3D、图生3D、八视图生3D、单几何生成（白模）功能。	• 文生3D • 图生3D	3

HY-3D-Express	hy-3d-express	采用混元生3D 极速版模型，可将生成模型时间缩短至1分30秒内，可在较短时间内生成3D 模型文件。	• 文生3D • 图生3D	1

能力说明

深度思考

模型在生成最终回答前，先进行内部思维链（Chain-of-Thought）推理，通过逐步分析和拆解问题，提升复杂任务（如数学、逻辑推理、代码生成等）的回答准确性。

联网搜索

模型支持在推理过程中访问互联网，检索并整合实时信息，从而为时效性问题（如新闻、天气、最新数据等）提供更准确的回答。

结构化输出

模型支持按照指定的格式（如 JSON Schema）输出结构化数据，便于下游程序直接解析和使用，适用于信息抽取、数据填充、API 响应构建等场景。

Function Calling

模型支持函数调用能力，可在推理过程中根据用户意图自动识别并触发预定义的外部工具或 API，实现查询数据库、调用第三方服务等扩展操作。

Cache 缓存

模型 Cache 缓存能力可复用历史请求中的上下文计算结果，减少重复计算开销，从而提升响应速度并降低调用成本。