API MODEL MATRIX · 核实于 2026-06-04

第三方 API 模型接入清单

当前工作流中通过 API 方式接入的全部第三方模型 —— 涵盖文本推理、图像生成、语音合成与数字人/视频四类。每一项标注模型名称、申请入口、能力价值与官方核实价格，便于横向对照与选型。

14 接入模型 / 服务
4 能力类别
9 厂商 / 平台
8/9 价格官方已核实

01 文本 LLM · 推理与对话

4 项

Claude（Anthropic）

`Opus 4.8 · Sonnet 4.6`

官方直连

● 主力

当前**推理质量天花板**，Agentic 编码与复杂多步任务最稳，工具调用可靠。Sonnet 享 1M 长上下文。

推理最强 · `Agentic 编码` · `长上下文 1M` · 工具调用稳

场景 Claude Code 核心引擎、复杂编码与多步推理

$5 / $25

Opus 4.8 输入 / 输出（每百万 token）
Sonnet 4.6 $3 / $15；缓存读取 0.1×；Batch 5 折

`console.anthropic.com` ↗

OpenRouter

`LLM 聚合路由`

300+ 模型统一网关

● 主力

一个 Key **统一访问 300+ 跨厂商模型**，自动故障转移与负载均衡，切换零成本，不被单一厂商锁定。

`300+ 模型` · `一个 Key` · 自动容灾 · 不锁厂商

场景 多模型统一调度、对话与内容生成的中转层

原价透传 不加价

充值收 5.5% 手续费；BYOK 每月前 100 万次免费、之后 5%

`openrouter.ai` ↗

Gemini Flash（Google）

`2.5 / 3.x Flash`

Google AI Studio

● 主力

高性价比多模态模型，**原生支持文件上传与视频分析**，超大上下文。适合大批量、低延迟的理解类任务。

多模态 · 视频分析 · 超大上下文 · 便宜

场景 视频内容分析、大批量文本理解与分类

$0.30 / $2.50

2.5 Flash 输入 / 输出（每百万 token）
3.5 Flash $1.50 / $9；Flash-Lite 低至 $0.10 / $0.40

`aistudio.google.com` ↗

Qwen3（SiliconFlow）

`Qwen3.5 系列`

硅基流动 · 国内直连

● 主力

国内直连的**开源 Qwen3 托管推理**，单价极低、低延迟，适合高频小任务。新账号含免费额度。

国内直连 · 单价极低 · 低延迟 · 开源模型

场景 意图分类、关键词抽取等高频轻量任务

¥0.40 起 / 百万 token

输入 ¥0.40–1.20（0–128k 档）
输出 ¥3.20–7.20；长上下文档约 2.5–4×

`siliconflow.cn` ↗

02 图像生成 · 文生图与编辑

6 项

Gemini 3 Pro Image

`gemini-3-pro-image-preview`

Google · nano-banana-pro

● 主力

原生 4K + 最高保真度，多参考图严格一致性，中文文字渲染最佳。40s 出 4K 不需 upscale。

`原生 4K` · 多参考图一致 · 中文文字准 · 色彩还原强

场景 商业级成片、4K 电商产品图、隐形模特

≈ $0.13 / 张（1–2K）

4K 约 $0.24/张；输入 $2、图像输出 $120（每百万 token）；Batch 半价

`aistudio.google.com` ↗

Gemini 3.1 Flash Image

`gemini-3.1-flash-image-preview`

Google · nano-banana 2

● 主力

1024 原生分辨率，**快且便宜**，中文渲染好。Pro 版的轻量替身，用于不追 4K 的高频出图。

快 · 便宜 · 中文好 · `1024 原生`

场景 中文 slogan 配图、内容快稿迭代

≈ $0.045 / 张起

$0.045（0.5K）→ $0.151（4K）按分辨率
输入 $0.50、图像输出 $60（每百万 token）

`aistudio.google.com` ↗

gpt-image-2（OpenAI）

`gpt-image-2`

OpenAI 官方直连

● 主力

指令跟随与图像编辑最强；保留原图主体换场景、重建光影逻辑。edits 最多 16 张参考图。

编辑最强 · 产品一致性 · 光影逻辑 · `16 张参考图`

场景 产品换场景、风格迁移、角色跨镜一致性

≈ $0.05 / 张（1024² 中）

低 $0.006 · 高 $0.21；图像输入 $8、输出 $30（每百万 token）；Batch 5 折

`platform.openai.com` ↗

即梦 Seedream 4.0

`doubao-seedream-4-0`

火山方舟 Ark · 国内直连

● 可用

中文理解与文字渲染强，多图融合（最多 10 张）、序列图（最多 15 张）。国内直连、单张成本最低。

中文准 · `多图融合 10` · `序列图 15` · 国内最便宜

场景 中文商业素材、批量风格统一配图

¥0.20 / 张

按输出图片计费（RMB）
商业素材务必走 API，不用 Web UI

`console.volcengine.com/ark` ↗

fal.ai

`FLUX / Seedream / Nano Banana`

多模型聚合平台

● 备用 · 余额耗尽

一个 API **聚合多家开源/托管图像视频模型**（含 Recraft 矢量化），按产出计费，排队与服务端错误不收费。

多模型聚合 · 矢量化 · 按产出计费 · 错误不收费

场景 跨模型快速试验、矢量 logo、备用通道

$0.03 / 张起

Seedream V4 $0.03 · Nano Banana $0.0398 · FLUX Kontext Pro $0.04/张

`fal.ai/pricing` ↗

可灵 Kling

`Kling · Kolors 可图`

快手可灵开放平台

● 暂停 · 无余额

视频生成业界领先（运动幅度 / 时长），图像生成（可图 Kolors）中英文字渲染强，面向 B2B 集成。

视频运动强 · 长时长 · 中英文字准 · B2B

场景 高动态视频生成（图像为辅）

$1 = 66 积分

≈$0.015/积分；视频 9–16 积分/秒
图像单价 **未公开核实**（站点拦截抓取）

`klingai.com/global/dev` ↗

03 语音 TTS · 合成与克隆

2 项

MiniMax TTS

`speech-2.8-hd`

MiniMax · 国内直连

● 主力

高拟真、情感丰富的中文 TTS，当前最高质量档。支持音色克隆、timber 混音（中文音色说地道英文）与跨语种。

最高拟真 · 音色克隆 · 混音 · 跨语种

场景 微信语音回复、视频号口播、数字人配音

¥3.5 / 万字符

turbo 档 **¥2/万字符**；按输入字符计，1 汉字 = 2 字符

`platform.minimaxi.com` ↗

Qwen3-TTS（阿里云百炼）

`qwen3-tts-flash · cosyvoice-v3-plus`

DashScope · 国内直连

● 迁移中

17 种方言 + 10 种外语，音色克隆 1 年有效（MiniMax 仅 7 天），自然语言情绪控制。单价低于 MiniMax。

`17 方言` · `克隆 1 年` · 情绪控制 · 更便宜

场景 微信语音（已上线）、视频号口播替换中

¥0.115 / 万字符

qwen3-tts-flash 国内价；音色复刻 $0.01/个

`bailian.console.aliyun.com` ↗

04 数字人 / 视频 · 生成

2 项

石榴数字人

`16ai · createByAudioFile`

向量方程科技

● 可用

一站式数字人：声音合成 + 形象训练（照片/视频建模）+ 音频驱动视频生成。覆盖广告、医疗、教育等行业。

一站式 · 音频驱动 · 照片建模 · 多行业

场景 数字人口播视频（TTS 出音 → 驱动形象）

¥6 / 分钟

数字人视频生成 **¥6 / 分钟**

`api.16ai.vip` ↗

即梦视频 Seedance

`doubao-seedance-1.0-pro / fast / lite`

火山方舟 Ark · 国内直连

● 可用

文生 / 图生视频。1.0-pro 高画质，pro-fast 提速 3 倍降价 72% 量产，lite 入门档。国内直连。

`文生/图生视频` · `高画质 pro` · `量产 fast` · 国内直连

场景 短剧、水墨漫剧、视频号批量出片

¥3.67 / 5s 1080p

1.0-pro ¥15/百万 token；pro-fast ≈¥1.03/5s；离线 batch 5 折

`console.volcengine.com/ark` ↗

● 主力 / 可用 —— 当前在用　● 迁移中 —— 逐步替换　● 备用 / 暂停 —— 余额或额度问题

价格说明 · 上表价格于 2026-06-04 核实自各厂商官方定价页，单位与币种以原页面为准（USD / RMB 混用，已标注）。除可灵图像单价（站点拦截抓取）外，其余均锚定官方页面。
价格随官方政策动态调整，实际计费以下单时官网为准。缓存、Batch、长上下文等折扣详见各官方文档。