



✦如何引入多模态大模型，丰富你的产品体验

白宦成

2025 年 9 月 20 日



如何引入多模态大模型，丰富你的产品体验

- 白宦成
 - 阶跃星辰开放平台产品经理
 - 开发者
 - 曾就职于字节跳动\腾讯等企业



关于阶跃星辰

阶跃星辰
是一家 AGI 创业公司

以实现通用人工智能为目标
专注研发 多模态基础大模型
基于自研基础大模型打造新一代 “人工智能+” 应用

成立时间
2023 年 4 月

成立地点
上海市徐汇区

员工数量
400+

研发人员占比
80%+

极高人才密度，聚集领先的计算机视觉、自然语言处理、分布式系统团队：SIGKDD 最佳应用论文奖、CVPR 最佳论文奖、物体识别与检测国际冠军、未来科技大奖、论文引用量超过 40 万



阶跃星辰 Step 系列通用大模型矩阵

14 个月发布 27 款模型，综合全面，多模领先

语言

Step-3

开源

多模态推理 | 推理效率
2025. 07

Step-2

API

新一代万亿MoE 性能 | 体感 | Planning
2024. 03

Step-1

API

千亿经典版
2024. 03

Step-R

开放

RL 范式
思维链 | 复杂推理
2025. 01 Step-R-mini
2025. 04 Step-R1-V-mini

多模态

图像理解

Step-1V

API

静态图像深度理解 | 上下文 8-32K
2024. 03

图像生成

Step-1X

API

强大的图像生成 | 原生中文支持
优秀美学 | 多种大小
2025. 04 Step-1X-Edit
2025. 05 Step-1X-3D

语音

Step-Audio

开源

千亿参数 | 生成理解一体化
超低延时 | 自然高质 | 双商在线
2025. 02

视频理解

Step-1.5V

API

真·视频理解 | 语义理解深刻
2024. 07

视频生成

Step-Video

开源

指令遵循 | 大幅度动作 | 中国风
2025. 02 Step-Video-T2V
2025. 03 Step-Video-T12V

音乐

音乐模型

开源

音乐生成 | 写词编曲 | 模仿秀
2025. 05



大模型很多,那什么是多模态大模型?



原来，这些都是多模态大模型的用法

原来 AI 也能帮我选西瓜？甚至还挺甜！



根据编号，我会推荐以下几个西瓜：

- 👑 1. 5号西瓜：颜色较深，花纹清晰，整体看起来比较成熟。
- 2. 9号西瓜：颜色均匀，花纹明显，看起来成熟度较好。
- 3. 10号西瓜：颜色深，底部也比较均匀，没有明显的缺陷。

这几个西瓜看起来成熟度比较高，可能会比较甜。可以再根据重量和敲击声音进一步确认。

Dea.moon.com



太长的课程看不动,要不让 AI 先帮我整理个重点?

从概念到理解:多模态大模型



什么是多模态？

- 多模态是指能够处理和理解多种不同类型数据（即“模态”）的能力,这些数据包括但不限于文本、图像、音频和视频。
- 其核心价值在于**实现模态间的深度融合与理解,而非简单的信息叠加**。
- 例如, 一个多模态模型在处理一张包含文字和图像的图表时, 不仅能识别图表的内容, 还能理解文字对图表的解释, 从而获得更全面的洞察。

Type

文本 (Text)

图像 (Image)

音频 (Audio)

视频 (Video)

Example

最基础的模态, 包括自然语言、代码等。

静态视觉信息, 如照片、图表、绘画等。

声音信息, 如语音、音乐、环境音等。

动态视觉信息, 是图像和音频的组合, 包含时间维度。



多模态大模型 VS 文本模态大模型

• 根据西瓜的编号,给我选瓜



根据编号, 我会推荐以下几个西瓜:

- 👑 1. **5号西瓜**: 颜色较深, 花纹清晰, 整体看起来比较成熟。
- 2. **9号西瓜**: 颜色均匀, 花纹明显, 看起来成熟度较好。
- 3. **10号西瓜**: 颜色深, 底部也比较均匀, 没有明显的缺陷。

这几个西瓜看起来成熟度比较高, 可能会比较甜。可以再根据重量和敲击声音进一步确认。

Des.moon.com

VS

我面前有10个西瓜,其中一号瓜的花纹是....二号瓜的花纹是...,请你帮我选一个瓜



模态转换意味着信息的压缩和损耗

- ASR 是一个很常见的语音转文本模型,其效果为将音频模态转换为文本模态.
- 转换的过程中仅保留的语音当中的语音文字信息,丢失了语音当中的情绪等信息.

春江潮水连海平，海上明月共潮生。滟滟随波千万里，何处春江无月明，江流宛转绕芳甸，月照花林皆似霰。空中流霜不觉飞，汀上白沙看不见。江天一色无纤尘，皎皎空中孤月轮。江畔何人初见月，江月何年初照人。人生代代无穷已，江月年年望相似。不知江月待何人，但见长江送流水。



主流的多模态大模型

模型	文本理解与生成	图像理解与分析	图像生成	音频处理	核心优势	是多模态大模型？
Google Gemini 2.0	强大，支持多种语言和代码	原生多模态，可进行复杂跨模态推理	不支持	能够理解和生成语音	原生多模态架构，跨模态推理能力强	✓
OpenAI GPT-4o	卓越，可进行长篇复杂写作	优秀，能处理复杂视觉任务	支持	实时语音交互能力突出	交互速度快，响应自然，适合实时应用	✓
Anthropic Claude 3	优秀，在长文本处理方面表现出色	强大，尤其擅长处理图表和文档	不支持	具备一定能力，但侧重文本和图像	安全性和可控性高，视觉理解能力强	✓
DeepSeek V3	优秀，擅长中文和代码处理	原生不支持	不支持	原生不支持	稀疏 MoE 架构，高性价比，中文能力强	✗
Step-3	卓越，在长文本和复杂推理上表现突出	原生不支持	不支持	原生不支持	多模态推理能力强，支持长文本和复杂任务	✓
Nano Banana(Gemini 2.5)	强大，支持多种语言和代码	原生多模态，可进行复杂跨模态推理	支持	能够理解和生成语音	原生多模态架构，跨模态推理能力强	✓



别人都在用多模态大模型做什么？



多模态大模型的应用场景：图片模态 + 文字模态

OPPO – 一键问屏, 用户按下 AI 键, 直接询问大模型, 图片当中的东西是什么?



多模态大模型的应用场景：图片模态 + 文字模态

OPPO – 一键问屏, 用户按下 AI 键, 直接询问大模型, 图片当中的东西是什么?



多模态大模型的应用场景：视频模态 + 文字模态

- 妙兜智能助手 – 通过视频,理解你在做什么,并替你操作



多模态大模型的应用场景：语音模态+ 文字模态

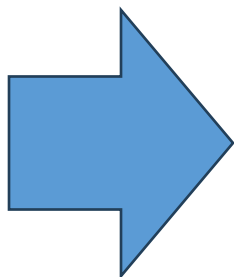
- System Prompt + 原生的语音理解生成能力.



多模态大模型的应用场景：图片模态的理解生成一体化

- Step-1X 图片编辑和理解能力

需求：请把车子外面的那个女人移除掉，其他东西都要保留，再重新输出图片



如何在你的产品当中加入多模态的能力？



痛点分析和价值提炼

- 不要为了多模态而多模态。
 - 它能解决什么具体问题？
 - 在内容管理系统中，用户想通过一张截图快速找到相关的文档；
 - 在社交产品中，用户想用语音来发送带有表情的文本消息。



低成本试错,验证需求的有效性

- 在快速验证想法或产品初期，直接调用云服务商（如 OpenAI 的 GPT-4o、Google Cloud AI 的 API 、 阶跃星辰的 Step 系列模型）是最快的路径。
 - 优点是部署简单，性能稳定；
 - 缺点是成本高，定制性差。
- 等业务稳定调用,验证价值之后,再评估考虑要不要自部署模型.



推荐一个快速开发的 Python SDK

- Pydantic AI: 框架本身抽象不错, 也适合 Python 开发者, 配合 LogFire , 可以非常方便的进行调用环节的 Debug.
- 要么, 就用 OpenAI 官方 SDK , 要么就用 Pydantic AI



Pydantic AI

GenAI Agent Framework, the Pydantic way



一些高级技巧



如何优化多模态大模型的指令遵循？

- 引入 Few-shot 来优化模型的输出结果；
- 优先加入正向 Case,尽量不加负向Case(负向注意力也是注意力,会影响模型效果)
- 在 System Prompt 中加入多模态内容,辅助模型理解.
- 对模型做 Finetune,对模型效果进行优化



如何优化多模态大模型的推理速度

- 构建固定的 System Prompt,使用 Prompt Cache ,减少推理的成本和耗时,提升整体的速度
- 优化输出格式,减少输出的 Token 数(比如用 Token 量更少的格式)
- 使用指令遵循更强的模型,从而减少 System Prompt 的 Token 数



如何应对多模态模型的“幻觉”问题，确保输出的准确性。

- Temperature 调低,让模型输出更大概率正确内容
- 数据非常重要: 干净的数据,会极大影响模型的回复的效果



阶跃星辰 繁星计划

- 阶跃星辰「繁星计划」是你的快速通道！我们正在寻找**多模态大模型领域的初创企业与优秀独立开发者/团队**，与我们一同打造行业多模态标杆案例，帮助开发者获得更大的影响力。
- 只要你对探索多模态大模型应用的无限可能性充满激情，那么你就是我们寻找的理想人选。加入我们，将有机会加速你的创新之旅，成为下一个“胃之书”！
- **你可以得到什么？**
 - 总价值4万元的tokens免费赠送
 - 线上线下曝光宣发资源（包括公众号、AI自媒体、线下大会等）。
 - 各行业合作伙伴资源牵线对接。
 - 结合项目背景，提供快速技术响应。
 - 孵化器场地免费工位支持。
 - 优先参与最新模型和功能试用。
 - RPM/TPM升级至VIP服务。
 - 优秀的项目和团队，将有机会获得阶跃星辰的直接投资，以加速项目落地和发展。



阶跃星辰 繁星计划



好事发生

☉ 类型: **打卡类**

☞ 一句话介绍:
记录并珍藏每个好消息

NiceBB

☉ 类型: **娱乐类**

☞ 使用模型: **Step-10-turbo-vision**

☞ 一句话介绍:
陪伴孩子度过美好童年

小映宝

☉ 类型: **医疗类**

☞ 使用模型: **Step-1V Step-1**

☞ 一句话介绍:
温暖陪伴, 专业呵护癌症患者/家属的 AI助手

歌词爆改机

☉ 类型: **音乐类**

☞ 使用模型: **Step-2 Step-1X**

☞ 一句话介绍:
生活乐子多, 不如改成歌

来萌LIMO

☉ 类型: **心理类**

☞ 使用模型: **Step-1**

☞ 一句话介绍:
万物岛小精灵LIMO们通过树洞与地球链接的桥梁

写本回忆录!!

小鹿光年

☉ 类型: **工具类**

☞ 一句话介绍:
让回忆成为跨越时空的拥抱

知潜 KnowFuture

☉ 类型: **招聘求职类**

☞ 使用模型: **Step-1V**

☞ 一句话介绍:
大模型驱动的人才平台

心光 - AI 生活记录伴侣

☉ 类型: **打卡类**

☞ 一句话介绍:
一款超精美的、会主动梳理和反馈的记录工具

deepeat

☉ 类型: **美食类**

☞ 使用模型: **Step-10-turbo-vision**

☞ 一句话介绍:
eat smarter, live better

MIKU 搜索

☉ 类型: **AI搜索类**

☞ 使用模型: **Step-2**

☞ 一句话介绍:
MIKU 搜索, AI 搜索 + 知识库的超级助手

每天拍云

☉ 类型: **娱乐类**

☞ 使用模型: **Step-1V**

☞ 一句话介绍:
天空收集者, 给每朵云都起了独特的名字

TheGame - 双人todo

☉ 类型: **打卡类**

☞ 使用模型: **Step-2**

☞ 一句话介绍:
双人协作任务清单工具

ReadPaper

☉ 类型: **科研类**

☞ 使用模型: **Step-1V**

☞ 一句话介绍:
基于 AI 科研大模型的工具平台和社区

食医伴侣

☉ 类型: **医疗类**

☞ 使用模型: **Step-10-turbo-vision**

☞ 一句话介绍:
糖友的私人营养师

职优越

☉ 类型: **招聘求职类**

☞ 使用模型: **Step-2**

☞ 一句话介绍:
第一个 AI 职业发展引擎

林间聊愈室

☉ 类型: **心理类**

☞ 使用模型: **Step-1V Step-2**

☞ 一句话介绍:
全球 Z 世代的 AI 情绪伙伴

灵魂摄取器

☉ 类型: **娱乐类**

☞ 使用模型: **Step-1X Step-1V**

☞ 一句话介绍:
万物皆有灵, 拍照发现你身边物品、宠物的灵魂!

APPLY U

☉ 类型: **招聘求职类**

☞ 使用模型: **Step-2**

☞ 一句话介绍:
AI 驱动的留学申请指导平台

物圆 TreLoop

☉ 类型: **电商类**

☞ 使用模型: **Step-1V**

☞ 一句话介绍:
AI 赋能的闲鱼分享与交易平台

陌生人闹钟

☉ 类型: **工具类**

☞ 使用模型: **Step-2**

☞ 一句话介绍:
让随机的陌生人叫你起床

赛博大舞台

☉ 类型: **娱乐类**

☞ 一句话介绍:
赛博大舞台, 有推你就来

胃之书

☉ 类型: **美食类**

☞ 使用模型: **Step-1V**

☞ 一句话介绍:
你的 AI 美食搭子

ChatPPT

☉ 类型: **效率类**

☞ 使用模型: **Step-1V**

☞ 一句话介绍:
免费【生成+编辑+美化+下载】的智能文档工具

HowCity

☉ 类型: **工具类**

☞ 一句话介绍:
Cross-Cultural AI Assistant

总结

- **多模态大有可为**：不要局限于文本，结合图像、音频和视频等模态，能为您的产品带来更具创新性和沉浸感的用户体验。
- **动手是最好的学习**：Python SDK 已经极大地简化了接入过程，从一个简单的“智能菜谱助手”开始，快速上手实践。
- **掌握技巧，事半功倍**：在实际项目中，请务必关注 RAG、工具调用、缓存与异步调用 等高级技巧，它们是确保应用稳定、高效、可扩展的关键。



Thanks!



白宦成

天津 河北



扫一扫上面的二维码图案，加我为朋友。

