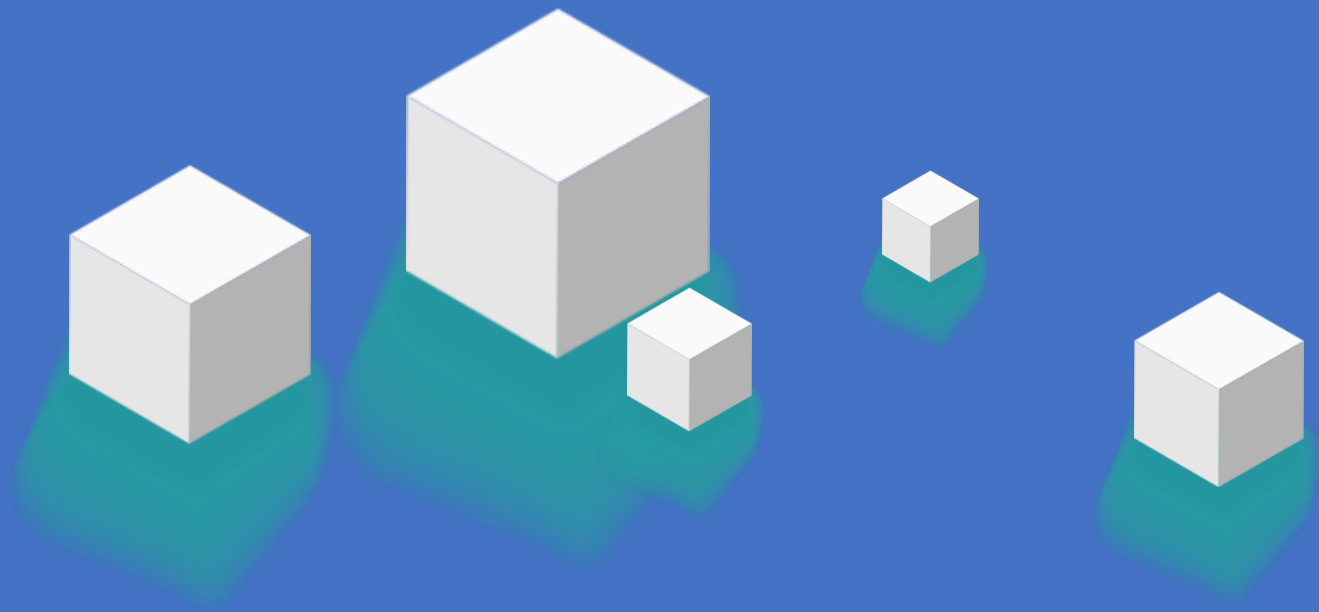
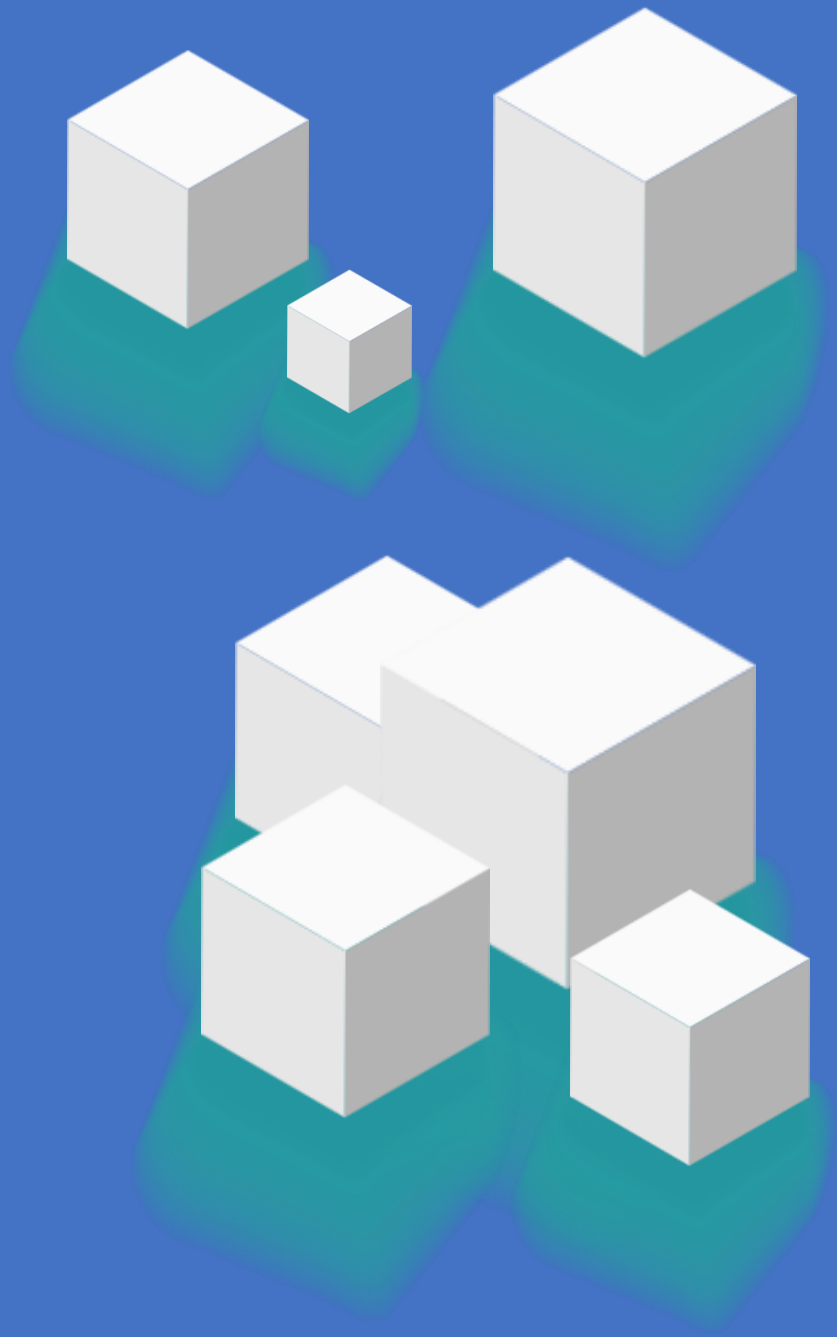
RAG多模态数据处理





>> 今天的学习目标

RAG多模态数据处理

- Gemini多模态处理

多模态输出（音频、视频、PPT）

Gemini API使用

- CASE：迪士尼 RAG 助手

Multimodal-Embedding 使用

Faiss索引构架

Query查询处理

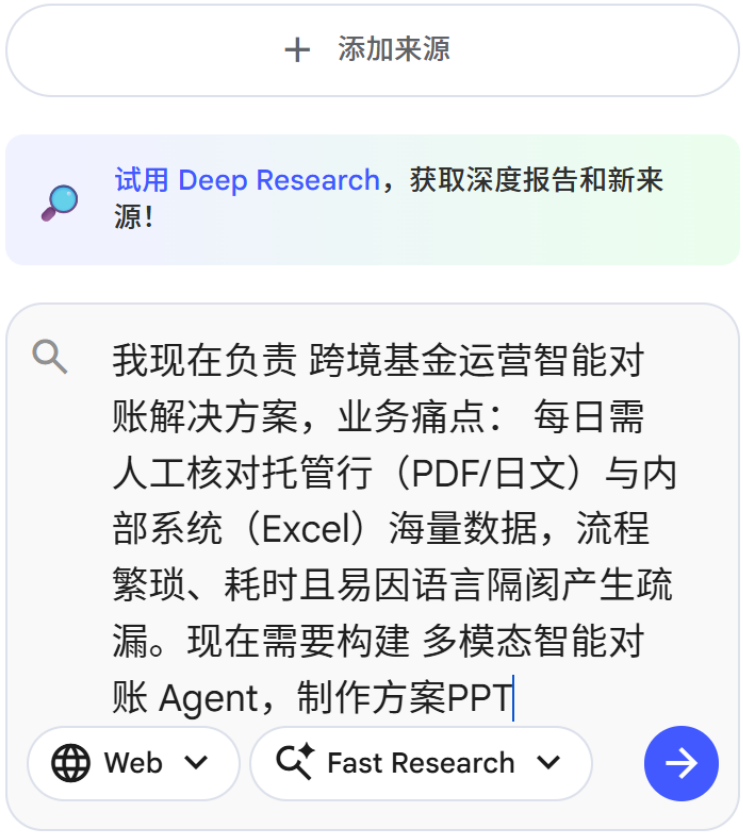
多模态统一向量空间

- 切片策略 (Chunking)

Gemini 多模态处理

CASE：快速了解某个领域的知识（Fast Research）

我现在负责 跨境基金运营智能对账解决方案，业务痛点： 每日需人工核对托管行（PDF/日文）与内部系统（Excel）海量数据，流程繁琐、耗时且易因语言隔阂产生疏漏。现在需要构建 多模态智能对账 Agent，制作方案PPT



CASE：快速了解某个领域的知识（生成音频）

自定义音频概览

格式

深入探究 ✓
两位主持人之间生动有趣的对话，旨在解读和关联来源中的主题

摘要
简短概要，旨在帮助您快速了解来源的核心思想

评论
对来源的专家评价，旨在提供建设性反馈，帮助您改进内容

辩论
两位主持人之间思维缜密的辩论，旨在阐明对来源的不同观点

选择语言

中文（简体）

时长

短 ✓ 默认

AI 主持人在本集节目中应着重于哪些方面？

提示示例
- 专注于特定来源（“仅介绍关于意大利的文章”）
- 专注于特定主题（“仅讨论小说的主角”）
- 针对特定受众（“向不熟悉生物学的人介绍生物学知识”）

CASE：快速了解某个领域的知识（生成视频）

CASE：快速了解某个领域的知识（生成PPT）

自定义演示文稿

格式

详细演示文稿 ✓

一整套包含全文和详情的演示文稿，非常适合通过邮件发送或单独阅读。

演示用幻灯片

简洁直观的幻灯片，附带要介绍的重点，为您的演讲提供全程支持。

选择语言

中文（简体）

时长

短 ✓ 默认

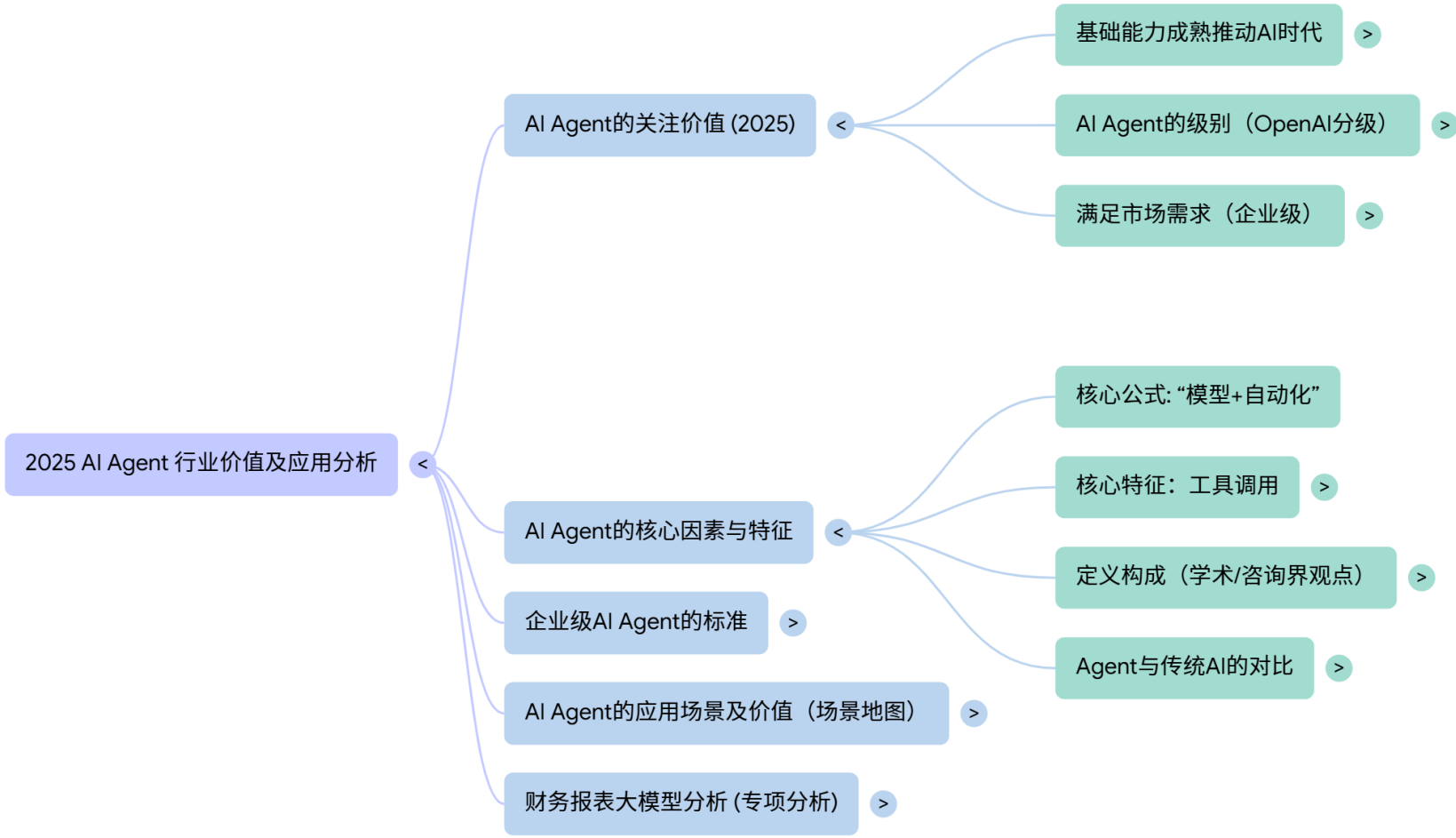
请描述您要创建的演示文稿

添加一份概略提纲，或指定受众、风格和重点：“为新手用户创建一套演示文稿，采用大胆活泼的风格，注重分步说明。”

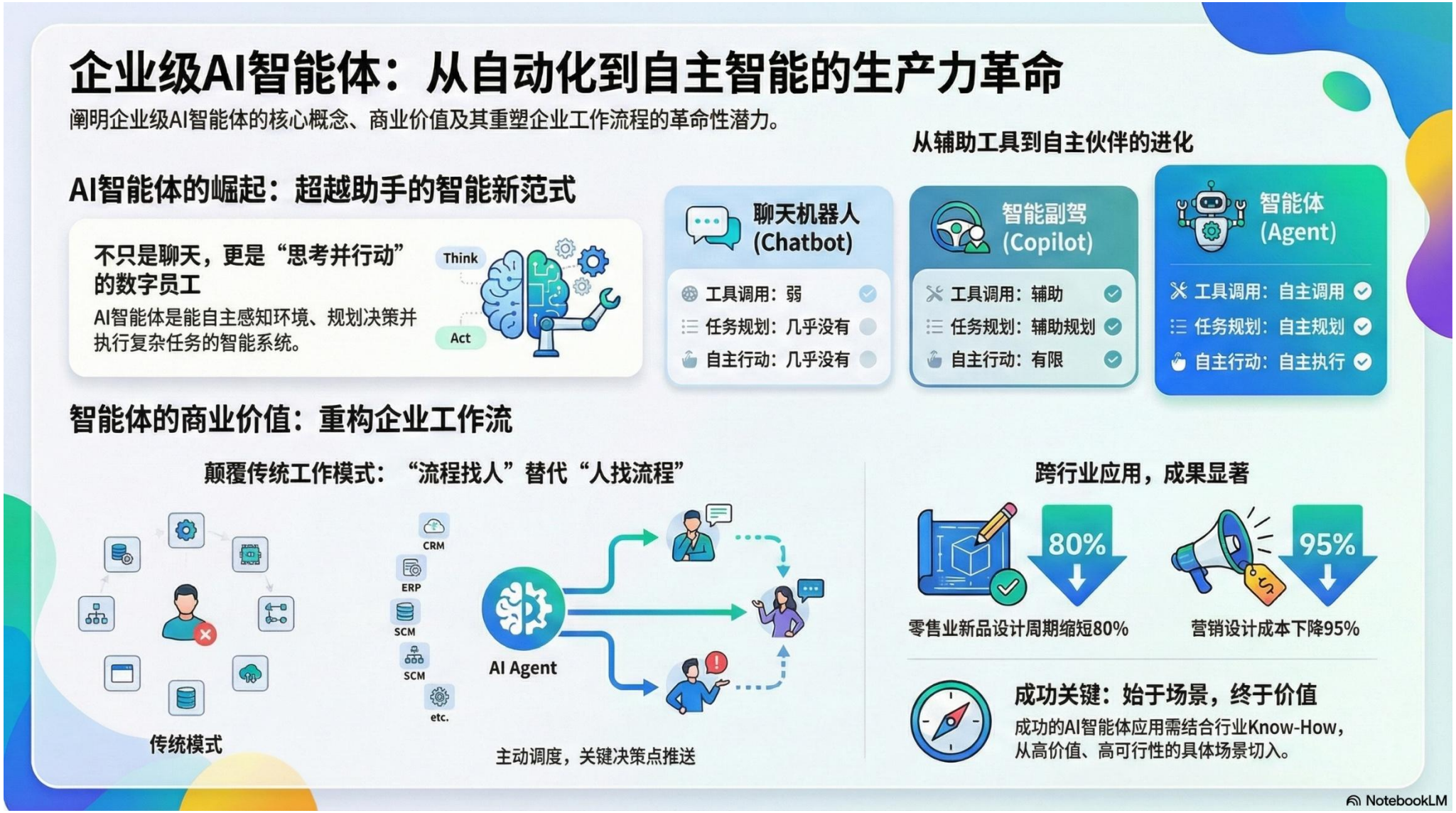
CASE：快速了解某个领域的知识（思维导图）

2025年AI Agent行业价值与应用分析

基于 7 个来源



CASE：快速了解某个领域的知识（信息图）



打卡：制作汇报材料

你有工作中需要讲解或者汇报的内容么？

比如项目的进度，系统方案的设计，项目管理的风险把控等，

再或者是某个新领域知识的学习，

某财务/人事制度的讲解等

结合自己的业务场景，制作材料：

- PPT演示
- 信息图
- 音频讲解（双人对话播客）
- 视频讲解

方案讲解中的配图

TO DO：讲解CNN的由来

我想要教授板书照片的样子：包含图表、箭头、方框和说明文字，从视觉上阐释核心思想。同时笔迹使用多种颜色，文字使用中文，下面是文章内容，你来帮我生成图片：

===

CNN的由来：

卷积神经网络（CNN）是人工神经网络的一种，是多层感知机（MLP）的一个变种模型，它是从生物学概念中演化而来的。

Hubel和Wiesel早期对猫的视觉皮层的研究中得知在视觉皮层存在一种细胞的复杂分布，这些细胞对于外界的输入局部是很敏感的，它们被称为“感受野”（细胞），它们以某种方法来覆盖整个视觉域。这些细胞就像一些滤波器一样，够更好地挖掘出自然图像中的目标的空间关系信息。

视觉皮层存在两类相关的细胞，S细胞（Simple Cell）和C（Complex Cell）细胞。S细胞在自身的感受野内最大限度地对图像中类似边缘模式的刺激做出响应，而C细胞具有更大的感受野，它可以对图像中产生刺激的模式的空间位置进行精准地定位。

S细胞（Simple Cell）：

作用是：找茬员。负责在感受野里找特定的线条（比如找竖线、找横线）

=> 相当于卷积层 (Convolution)

C细胞（Complex Cell）：

作用是：概括员。它不在乎线条具体在哪一毫米，只要这块区域里“有这个特征”就行。这就是为什么你的脸在照片左边还是右边，CNN 都能认出来（平移不变性）。

=> 相当于池化层 (Pooling)

方案讲解中的配图

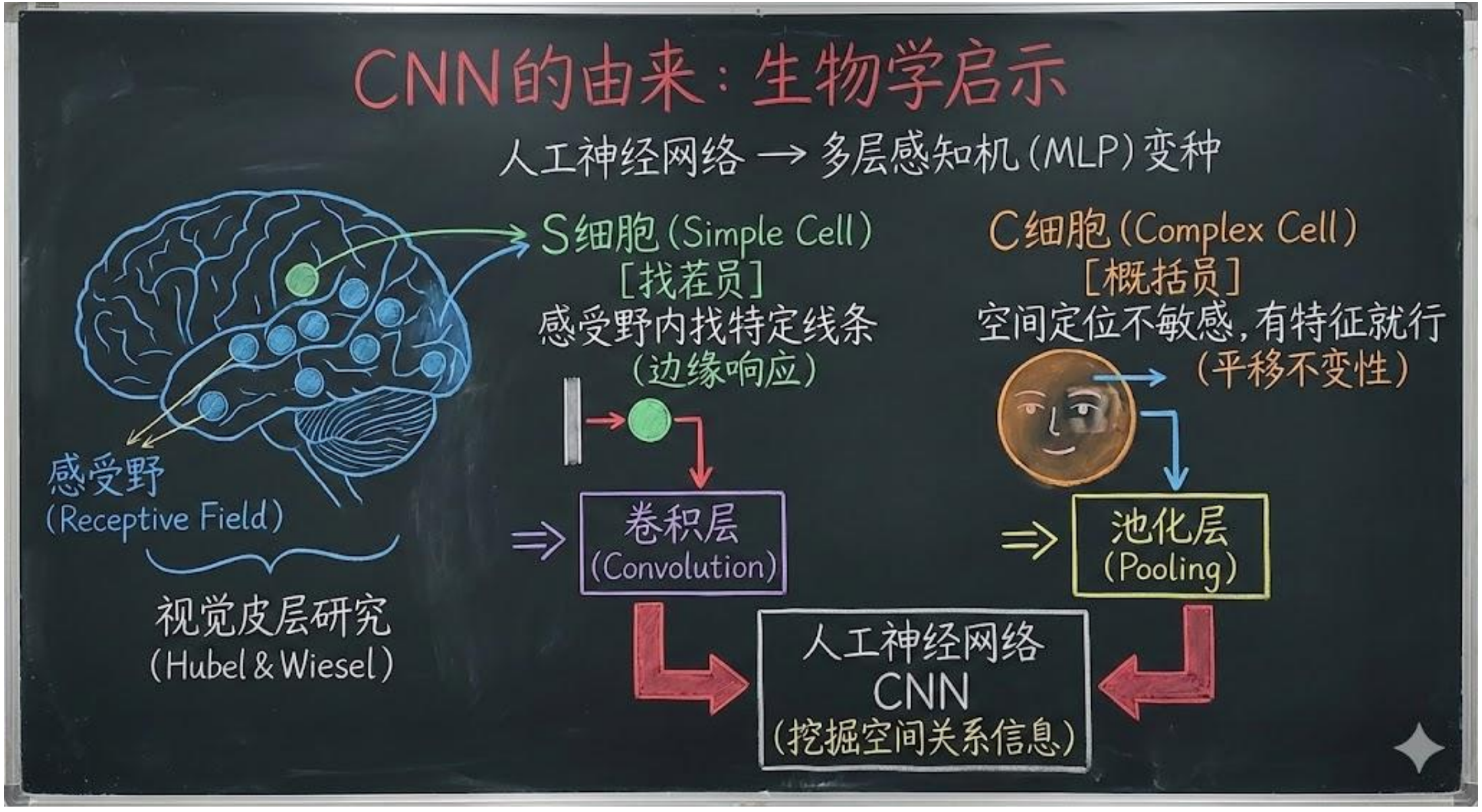
我来为您生成教授板书风格的图片，将CNN的由来和原理用清晰的板书形式展现出来。

我已经为您生成了4张教授板书风格的图片，展现了CNN的由来和原理。这些板书包含了您
箭头和方框等视觉元素来帮助理解。

您觉得这些板书图片如何？如果需要调整布局、颜色搭配或者内容的呈现方式，请告诉我！

除了使用即梦，还有哪些文生图好用的工具？

方案讲解中的配图


CNN的由来：生物学启示
人工神经网络 → 多层感知机（MLP）变种
S细胞（Simple Cell）
[找茬员]
感受野内找特定线条
（边缘响应）
C细胞（Complex Cell）
[概括员]
空间定位不敏感，有特征就行
（平移不变性）
感受野
（Receptive Field）
视觉皮层研究
（Hubel & Wiesel）
卷积层
（Convolution）
池化层
（Pooling）
人工神经网络
CNN
（挖掘空间关系信息）

Nano banana

Nano Banana 是 Google推出的最新一代 AI 图像生成与编辑模型

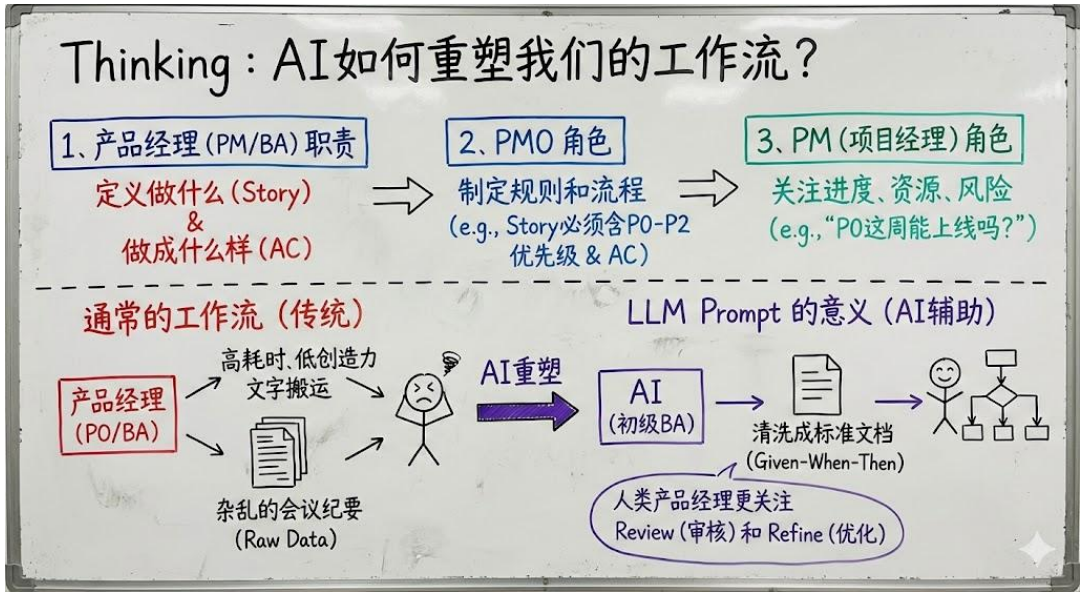
正式名称通常对应 Gemini 2.5 Flash Image 或 Gemini 3.0 Pro Image

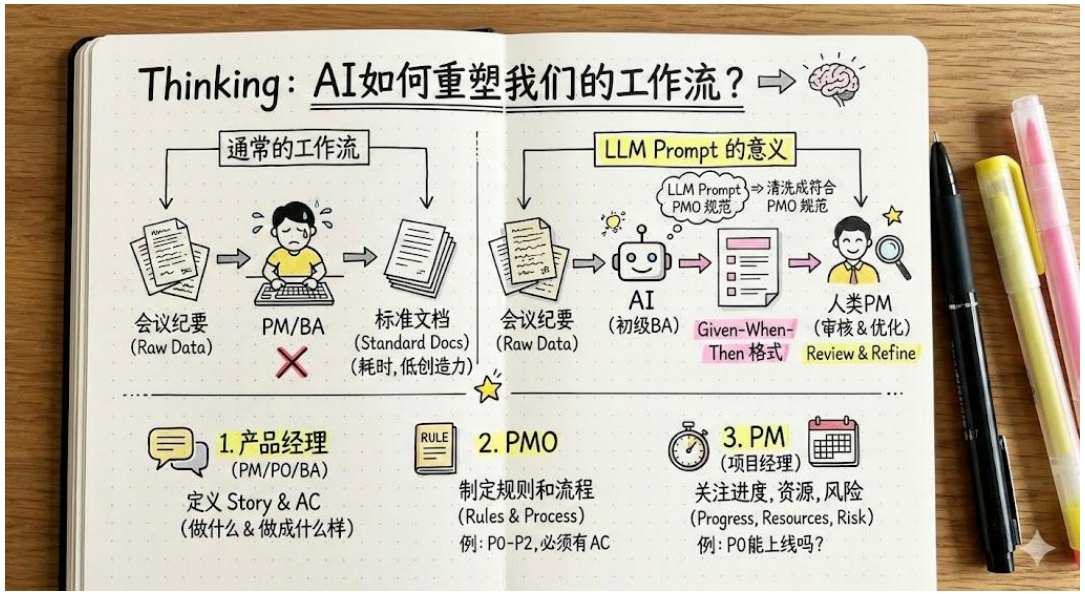
这个名字之所以在 AI 圈和社交媒体上非常火，是因为它最初在模型竞技场（LMArena）进行“盲测”时，使用了 "Nano Banana" 这个匿名代号，结果因为生成效果惊艳（尤其是写实感和文字渲染能力）而迅速出圈。

Thinking：为什么Nano Banana很强?

- 超强写实感：生成的照片级图像非常逼真，皮肤纹理、光影处理都达到了新高度。
- 文字渲染能力：以前的 AI 很难在图片里写对字（比如招牌、海报上的文字），Nano Banana 能比较准确地生成图片中的英文或其他文字。
- 一致性（Consistency）：它能更好地保持人物或物体的一致性。比如可以生成同一个角色在不同场景下的照片，脸不会变。
- 世界知识（World Knowledge）：结合了 Google 的搜索能力，它对现实世界的地标、物品理解更准确。

都有哪些典型的图像创作，可以用于日常的工作汇报中？





都有哪些典型的图像创作，可以用于日常的工作汇报中？

风格1：白板书

我想要白板书的样子：包含图表、箭头、方框和说明文字，从视觉上阐释核心思想。同时笔迹使用多种颜色，文字使用中文，下面是我的内容：{你的文字内容}

风格2：视觉笔记

一页打开的Moleskine笔记本，上面写满了精美的视觉笔记(Sketchnote)。背景是米黄色的点阵纸。使用黑色中性笔书写中文内容，并用黄色和粉色的马克笔进行高光标注。以下是我的内容：{你的文字内容}

风格3：未来科技风

一张高科技风格的未来主义数据分析大屏，全息投影界面(HUD)。深蓝色和黑色背景，带有霓虹青色和橙色的发光线条。画面结构清晰，包含模块化的数据方框、流程图箭头和发光的图表，图片中要用中文，根据下面的内容来生成：{你的文字内容}

风格4：3D黏土风

3D 轴测视图(Isometric view) 的概念图，C4D 或Blender 渲染风格，黏土材质(Claymorphism)。背景是柔和的哑光深灰色。场景构建：场景中心是几个立体的方块平台，上面漂浮着3D的中文字体“AI重塑工作流”，根据下面的内容来生成：{你的文字内容}

Gemini 的多模态处理能力

Gemini 的多模态处理能力：

把文本、图像、音频、视频、PDF、代码等不同形态的数据一次性喂给模型，它能在同一套神经网络里完成理解、推理、生成，无需先转文本。

- 原生统一架构

从预训练开始就把所有模态放在同一表征空间学习，而不是“先训大语言模型再外挂视觉/语音模块”，因此信息损耗更小、时序和细节保留更完整 。

- 端到端推理

同一组 Transformer 参数直接处理任意组合输入：

✓ 给一张 CT 影像 + 病历文本，能同步给出诊断建议；

✓ 给一段手写食谱视频，可直接翻译成可分享的数字菜谱 。

Gemini 的多模态处理能力

- 上下文规模

Gemini 3 Pro 支持 100 万 token 长窗口

可一次性读入 1 小时视频或 700 页 PDF，再输出 6.4 万 token 结构化报告 。

- 生成能力

不仅能“看”和“听”，还能“画”和“说”：

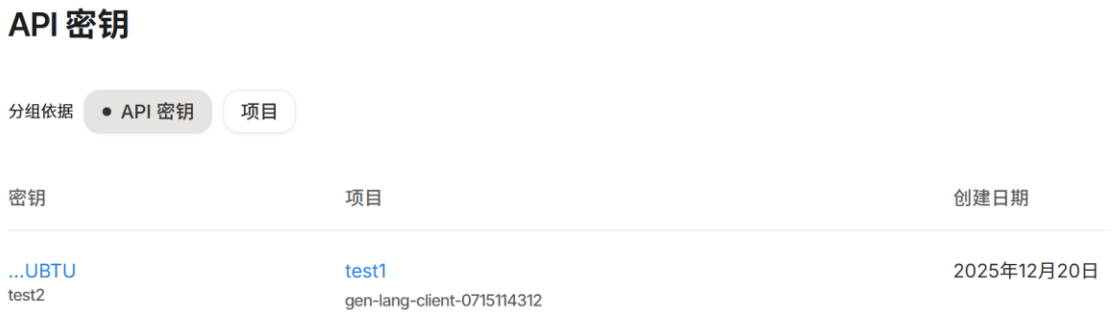
✓ 文本生成图像、编辑图像、褪色修复、3D 模型输出；

✓ 流式生成音频，实现边看边解说 。

CASE： Gemini 多模态处理

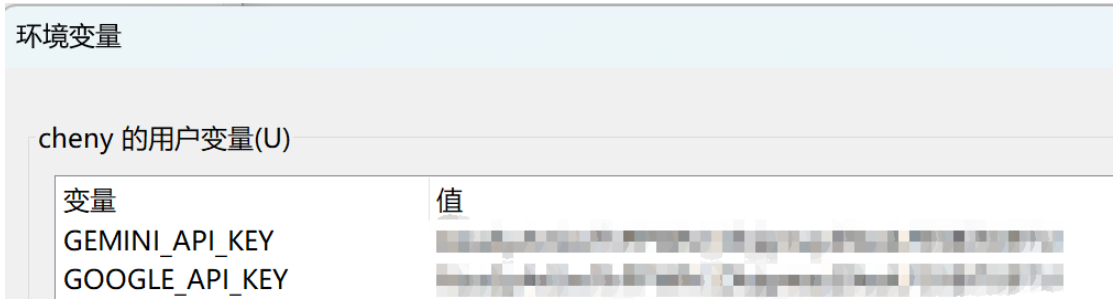
Step1， 申请 Gemini API KEY

https://aistudio.google.com/app/api-keys



Step2，添加环境变量

设置 GEMINI_API_KEY 和 GOOGLE_API_KEY



CASE： Gemini 多模态处理

```
from google import genai


client = genai.Client()
# 文字输出
response = client.models.generate_content(
    model="gemini-3-flash-preview",
    contents="用中文解释AI大模型是如何工作的",
)
print(response.text)
```

通俗地讲，AI大模型（如ChatGPT、Claude、文心一言等）就像是一个**“读过全人类几乎所有书的超级博学者”**。

要理解它是如何工作的，我们可以将其拆解为四个核心步骤：**喂养数据、理解上下文、预测下一个词、人工调教**。

---

1. 核心逻辑：概率预测（接龙游戏）
从本质上讲，AI大模型目前做的最核心的事只有一件：**根据上文，预测下一个字（或词）出现的概率。**

* **例子**：如果你给AI输入“床前明月”，它会根据在大规模语料中学到的知识，计算出下一个字是“光”的概率最高（比如99%）。
* **进化**：大模型之所以“大”，是因为它不仅能接唐诗，还能接代码、法律条文、甚至是复杂的逻辑推理。它把这种“接龙”做到了极致。

2. 核心架构：Transformer（注意力机制）
为什么现在的AI比以前聪明得多？主要是因为一种叫 **Transformer** 的技术架构。它引入了**“注意力机制”（Attention）**。
…

CASE： Gemini 多模态处理

```
from PIL import Image
# 图像理解
image = Image.open("dog_and_girl.jpeg")


# 注意：contents 变成了一个列表，里面同时放了图片对象和文字
response = client.models.generate_content(
    model="gemini-3-flash-preview",
    contents=[image, "帮我解释下这张照片"]
)


print(response.text)
```

CASE： Gemini 多模态处理

```
# 视频理解
import time

# 1. 上传视频文件
print("正在上传视频...")
video_file = client.files.upload(file="car.mp4") # 汽车剐蹭视频
print(f"上传成功: {video_file.name}")

# 2. 等待视频处理 (关键步骤！)
# 视频上传后，Google 需要几秒钟在云端进行转码。
while video_file.state.name == "PROCESSING":
    print("视频处理中，请稍候...")
    time.sleep(2)
    video_file = client.files.get(name=video_file.name)

if video_file.state.name == "FAILED":
    raise ValueError("视频处理失败")
print("视频就绪，开始推理...")

# 3. 多模态推理
response = client.models.generate_content(
    model="gemini-3-flash-preview",
    contents=[
        video_file,
        "详细描述视频里发生了什么？如果有对话，请把关键对话提取出来。"
    ]
)
print(response.text)
```

1-Gemini多模态处理.py

CASE：迪士尼RAG助手

CASE：迪士尼RAG助手

TO DO：搭建迪士尼RAG助手

为迪士尼构建一个7x24小时在线的AI客服助手：

- 自动化解答高频问题：如票务、入园须知、会员权益等，降低人工客服压力。
- 提供准确的回答：确保所有回答均来自官方知识库，避免信息错误或过时。
- 处理多模态查询：不仅能回答文本问题，还能理解并回应关于图片（如活动海报）的查询。

CASE：迪士尼RAG助手

Thinking：挑战都有哪些？

知识来源多样化：知识库包含多种格式的文档，如 PDF 格式的官方规定、Word 格式的内部FAQ、网页公告、以及包含大量图片和表格的活动介绍文件。

- 非结构化数据处理：如何有效提取并理解 PDF 和 Word 文档中的表格与图片信息，这是RAG成功的关键。
- 知识的有效组织：如何将海量、零散的知识点切片（Chunking）并建立索引，确保检索的准确性。
- 保证答案的有效性：如何让最终生成的答案严格基于检索到的内容，避免LLM出现幻觉。

CASE：迪士尼RAG助手

Thinking：技术选型是怎样的？

方案1：

- 文档处理库: PyMuPDF (处理PDF), python-docx (处理Word), pytesseract (OCR识别图片中的文字)。
- 文本Embedding模型: text-embedding-v4 (性能优秀，支持可变维度)。
- 图像Embedding模型: CLIP (由OpenAI开发，能同时理解图片和文本，是多模态RAG的核心)。
- 向量数据库/库: FAISS，作为向量检索引擎的核心，性能极高

注意：在生产环境中，可以使用Milvus, ChromaDB 或 Elasticsearch，他们提供了完整数据管理服务

- LLM: Qwen-turbo 用于最终答案的生成。
- 流程编排框架: 不依赖于LangChain，直接使用底层API

CASE：迪士尼RAG助手

Thinking：技术选型是怎样的？

方案2：

文本Embedding、图像Embedding、视频Embedding 都采用 Multimodal-Embedding

即多模态向量模型将文本、图像或视频转换成统一的1024维浮点数向量

- 跨模态检索：实现以文搜图、以图搜视频、以图搜图等跨模态的语义搜索。
- 语义相似度计算：在统一的向量空间中，衡量不同模态内容之间的语义相似性。
- 内容分类与聚类：基于内容的语义向量进行智能分组、打标和聚类分析。

Multimodal-Embedding 使用

模型名称	向量维度	文本长度限制	图片限制	视频片限制	单价（每千输入Token）
qwen2.5-vl-embedding	2048, 1024, 768, 512	32,000 Token	≤5MB,1张	≤50MB	图片/视频：0.0018元 文本：0.0007元
tongyi-embedding-vision-plus	1,152	1,024 Token	≤3MB,≤8张	≤10MB	0.0005元
tongyi-embedding-vision-flash	768	1,024 Token	≤3MB,≤8张	≤10MB	0.00015元
multimodal-embedding-v1	1,024	512 Token	≤3MB,1张	≤10MB	图片/视频：0.0009 元 文本：0.0007 元

所有模态（文本、图片、视频）生成的向量都位于同一语义空间，

可直接通过距离/余弦相似度等进行跨模态匹配与比较。

Multimodal-Embedding 使用（文本）

```
import dashscope
import json
from http import HTTPStatus

text = "上海迪士尼乐园门票分为一日票、两日票和特定日票三种类型。一日票可在购买时选定日期使用，价格根据季节浮动，平日成人票475元起"
input = [{'text': text}]
# 调用模型接口
resp = dashscope.MultiModalEmbedding.call(
    model="tongyi-embedding-vision-plus",
    input=input
)

if resp.status_code == HTTPStatus.OK:
    result = {
        "status_code": resp.status_code,
        "request_id": getattr(resp, "request_id", ""),
        "code": getattr(resp, "code", ""),
        "message": getattr(resp, "message", ""),
        "output": resp.output,
        "usage": resp.usage
    }
    print(json.dumps(result, ensure_ascii=False, indent=4))
```


```
{
    "status_code": 200,
    "request_id": "54b6b774-d1c5-4447-a126-e9322847babd",
    "code": "",
    "message": "",
    "output": {
        "embeddings": [
            {
                "embedding": [
                    0.0194854736328125,
                    0.02227783203125,
                    -0.04229736328125,
                    0.004268646240234375,
                    -0.0081329345703125,
                    ...
```


1-文本embedding.py

Multimodal-Embedding 使用（图片）

```
import dashscope
import base64
import json
from http import HTTPStatus
# 读取图片并转换为Base64
image_path = "./disney_knowledge_base/images/1-聚在一起说奇妙.jpg"
with open(image_path, "rb") as image_file:
    # 读取文件并转换为Base64
    base64_image = base64.b64encode(image_file.read()).decode('utf-8')
image_format = "jpg"  # 根据实际情况修改，比如jpg,bmp,png 等
image_data = f"{base64_image}"
input = [{'image': image_data}]
# 调用模型接口
resp = dashscope.MultiModalEmbedding.call(
    model="tongyi-embedding-vision-plus",
    input=input
)
```

```
if resp.status_code == HTTPStatus.OK:
    result = {
        "status_code": resp.status_code,
        "request_id": getattr(resp, "request_id", ""),
        "code": getattr(resp, "code", ""),
        "message": getattr(resp, "message", ""),
        "output": resp.output,
        "usage": resp.usage
    }
    print(json.dumps(result, ensure_ascii=False, indent=4))
```

```
{
    "status_code": 200,
    "request_id": "6baa87ee-52e3-41f4-ba91-06ad7a1c3662",
    "code": "",
    "message": "",
    "output": {
        "embeddings": [
            {
                "embedding": [
                    -0.0235595703125,
                    ...
```

2-图片embedding.py

Multimodal-Embedding 使用（视频）

```
# 多模态向量化模型目前仅支持以URL形式输入视频文件，暂不支持直接传入本地视频。
import dashscope
import json
from http import HTTPStatus
# 实际使用中请将url地址替换为您的视频url地址
video = "https://dataset-1255932437.cos.ap-nanjing.myqcloud.com/mp4/car.mp4"
input = [{'video': video}]
# 调用模型接口
resp = dashscope.MultiModalEmbedding.call(
    model="tongyi-embedding-vision-plus",
    input=input
)
if resp.status_code == HTTPStatus.OK:
    result = {
        "status_code": resp.status_code,
        "request_id": getattr(resp, "request_id", ""),
        "code": getattr(resp, "code", ""),
        "message": getattr(resp, "message", ""),
        "output": resp.output,
        "usage": resp.usage
    }
    print(json.dumps(result, ensure_ascii=False, indent=4))
```


```
{
    "status_code": 200,
    "request_id": "321fb55b-8374-43bc-9300-378355f57b61",
    "code": "",
    "message": "",
    "output": {
        "embeddings": [
            {
                "embedding": [
                    -0.0207366943359375,
                    0.02093505859375,
                    -0.03759765625,
                    -0.01044464111328125,
                    0.0093994140625,
                    ...
```


3-视频embedding.py

格式处理相关（.docx）

- 整体功能:

函数 parse_docx 读取一个 .docx 文件，遍历文件中的所有元素 (段落和表格)，并将它们提取成一个个独立的内容区块 (chunks)。

- 段落处理 (if element.tag.endswith('p')):

找出文档中的所有段落。

它会逐一提取每个段落内的纯文本内容，去除多余的空白，然后将其标记为 "type": "text" 并存储起来。

- 表格处理 (elif element.tag.endswith('tbl')):

专门处理文档中的“表格”。

它会将 Word 原生的表格格式转换为通用的 Markdown 格式。程序会先读取表头，然后逐行读取单元格数据，最后组合成 Markdown 表格字符串，并标记为 "type": "table"。

格式处理相关（.docx）

```
def parse_docx(file_path):
    doc = DocxDocument(file_path)
    content_chunks = []

    for element in doc.element.body:
        if element.tag.endswith('p'):  # 段落处理
            paragraph_text = ""
            for run in element.findall('.//w:t', {'w':
'http://schemas.openxmlformats.org/wordprocessingml/2006/main'}):
                paragraph_text += run.text if run.text else ""
            if paragraph_text.strip():
                content_chunks.append({"type": "text", "content":
paragraph_text.strip()})
        elif element.tag.endswith('tbl'):  # 表格处理
            # 转换为Markdown格式
            md_table = []
            table = [t for t in doc.tables if t._element is element][0]
            if table.rows:
                header = [cell.text.strip() for cell in table.rows[0].cells]
                md_table.append("| " + " | ".join(header) + " |")
                md_table.append("|" + "---|"*len(header))

                for row in table.rows[1:]:
                    row_data = [cell.text.strip() for cell in row.cells]
                    md_table.append("| " + " | ".join(row_data) + " |")

                table_content = "\n".join(md_table)
                if table_content.strip():
                    content_chunks.append({"type": "table", "content":
table_content})

    return content_chunks
```

disney_bot.py

格式处理相关（.docx）

- 整体功能:

函数 parse_docx 读取一个 .docx 文件，遍历文件中的所有元素 (段落和表格)，并将它们提取成一个个独立的内容区块 (chunks)。

- 段落处理 (if element.tag.endswith('p')):

找出文档中的所有段落。

它会逐一提取每个段落内的纯文本内容，去除多余的空白，然后将其标记为 "type": "text" 并存储起来。

- 表格处理 (elif element.tag.endswith('tbl')):

专门处理文档中的“表格”。

它会将 Word 原生的表格格式转换为通用的 Markdown 格式。程序会先读取表头，然后逐行读取单元格数据，最后组合成 Markdown 表格字符串，并标记为 "type": "table"。

格式处理相关（.pdf）

```
def parse_pdf(file_path, image_dir):
    doc = fitz.open(file_path)
    content_chunks = []

    for page_num, page in enumerate(doc):
        # 提取文本
        text = page.get_text("text")
        content_chunks.append({"type": "text", "content": text, "page":
page_num + 1})

        # 提取图片
        for img_index, img in enumerate(page.get_images(full=True)):
            xref = img[0]
            base_image = doc.extract_image(xref)
            image_bytes = base_image["image"]
            image_ext = base_image["ext"]
            image_path = os.path.join(image_dir,
f"{os.path.basename(file_path)}_p{page_num+1}_{img_index}.{image_ext}")

            with open(image_path, "wb") as f:
                f.write(image_bytes)
            content_chunks.append({"type": "image", "path": image_path,
"page": page_num + 1})

    return content_chunks
```

disney_bot.py

格式处理相关（.pdf）

- 整体功能:

parse_pdf 使用 fitz (PyMuPDF) 库打开并逐页读取 PDF 文档。它的核心目标是将 PDF 这种复合式文档，拆解成纯文本和独立的图片文件，以便 RAG 模型后续能分别处理和理解。

- 文本提取 (page.get_text("text")):

代码遍历每一页，使用 get_text() 方法抓取该页所有的纯文本内容。

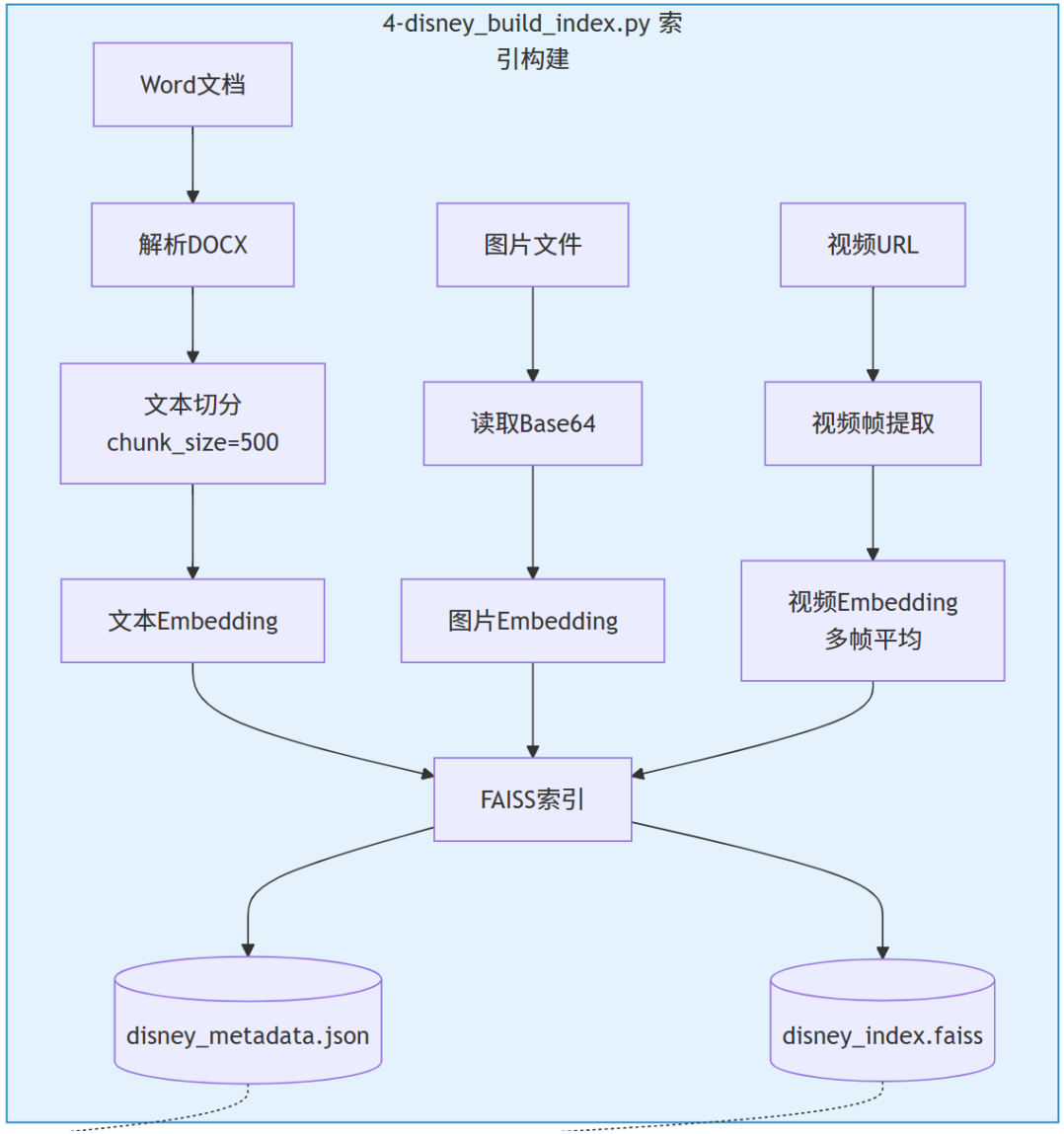
它将每页的文本保存为一个独立的区块 (chunk)，并附上页码。

- 图片提取 (page.get_images()):

侦测并提取页面中的所有嵌入图片。

它将每张图片的二进制数据读取出来，以唯一的文件名（包含原始文件名、页码和图片索引）保存到指定的 image_dir 目录下。同时，它会记录图片的存储路径。

Faiss索引构架



Step1，解析Word文档

parse_docx() - 遍历DOCX元素，提取段落文本和表格(转Markdown格式)

Step2，文本切分

split_text() - 按固定长度切分，chunk_size=500字符，overlap=50字符重叠

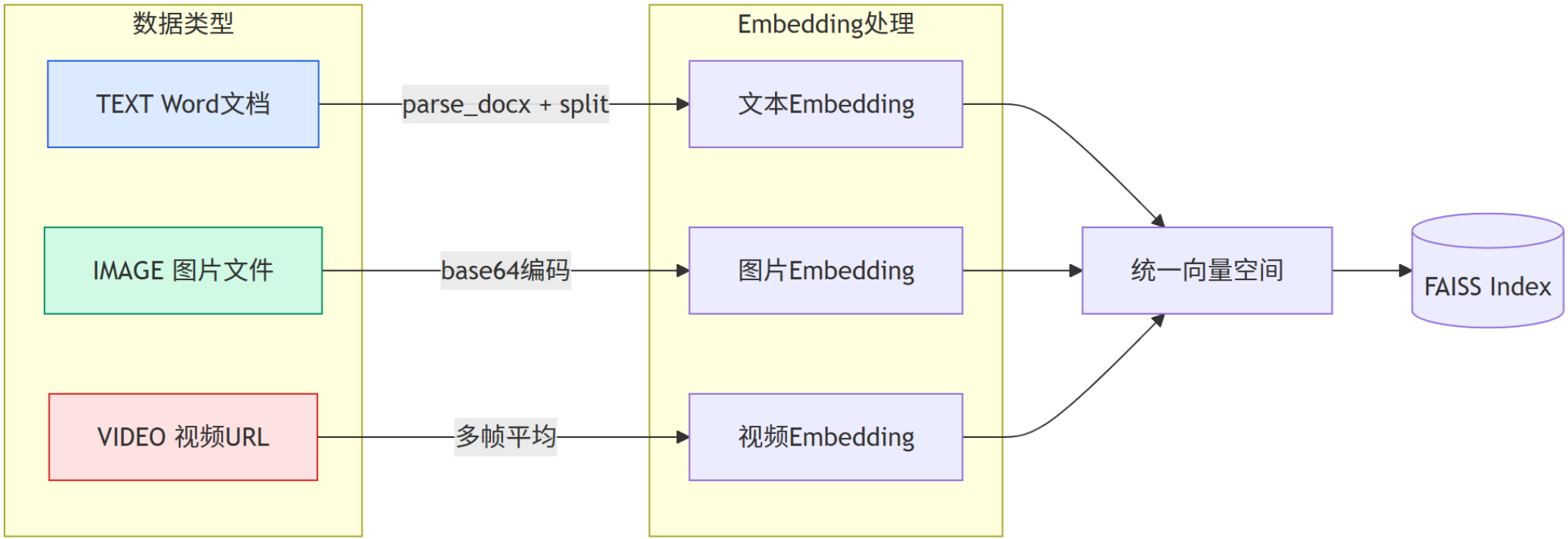
Step3，多模态Embedding

使用 tongyi-embedding-vision-plus 模型统一处理文本/图片/视频

- 文本: 直接编码
- 图片: Base64编码后发送
- 视频: 多帧提取后取平均向量

4-disney_build_index.py

Faiss索引构架



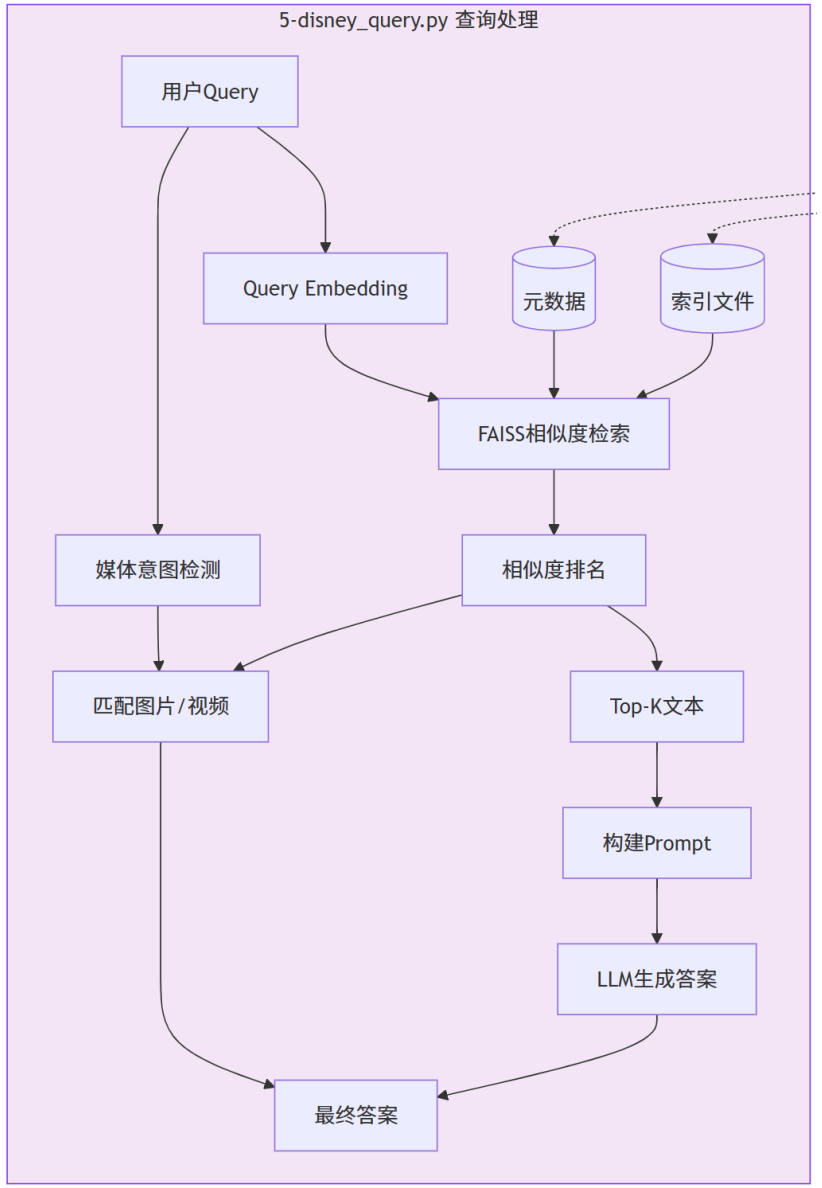
Step4，构建FAISS索引：IndexFlatL2 - L2距离(欧氏距离)精确搜索索引

Step5，持久化存储

索引 -> disney_index.faiss，元数据 -> disney_metadata.json

4-disney_build_index.py

Query查询处理



Step1，加载索引

load_index() - 从文件加载FAISS索引和元数据JSON

Step2，Query Embedding

将用户问题转为向量，与索引中的向量在同一空间

Step3，相似度检索

检索全部记录，按L2距离排序，转换为相似度: sim = 1/(1+distance)

Step4，媒体意图检测

关键词匹配: 图片["图片","海报","照片"...] | 视频["视频","录像","播放"...]

Step5，结果筛选

- 文本: 取Top-K (默认k=3)

- 图片/视频: 距离小于3.0的最近匹配

5-disney_query.py

Step6，LLM生成

构建Prompt (背景知识 + 用户问题)，调用 qwen-flash 生成答案

Query查询处理

```
// 文本类型
{
 "id": 0,
 "source": "退票政策.docx",
 "type": "text",
 "content": "退票内容..."
}
// 图片类型
{
 "id": 10,
 "source": "图片: poster.jpg",
 "type": "image",
 "path": "images/poster.jpg",
 "content": "[图片] poster.jpg"
}
```

```
// 视频类型
{
 "id": 15,
 "source": "视频: 汽车剐蹭",
 "type": "video",
 "url": "https://...",
 "description": "汽车剐蹭视频"
}
```

Metadata 元数据格式

Query查询处理

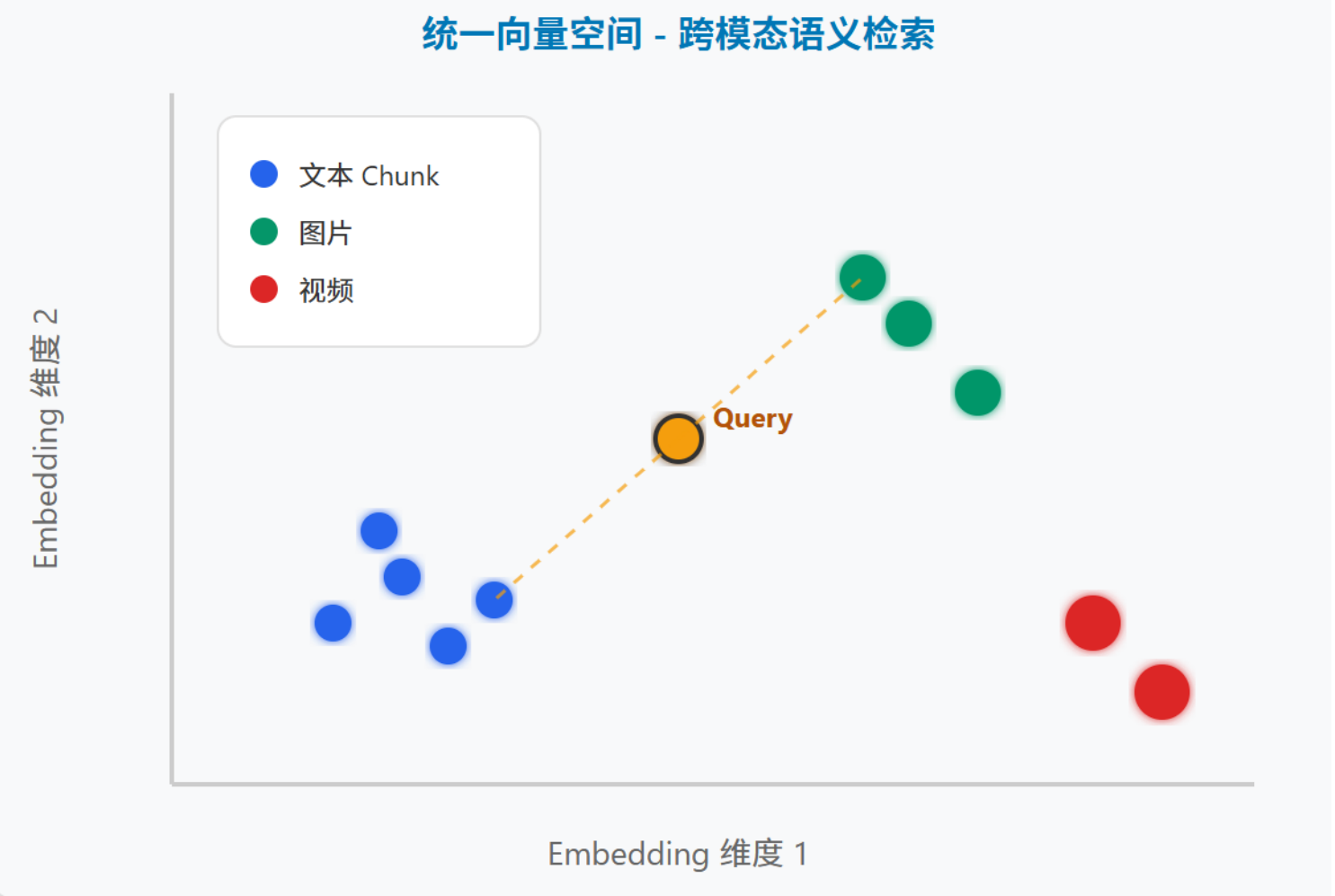
```
# 切分参数
CHUNK_SIZE = 500
CHUNK_OVERLAP = 50
# 媒体匹配阈值
MEDIA_DISTANCE_THRESHOLD = 3.0


# 关键词检测
IMAGE_KEYWORDS = [
 "图片", "海报", "照片",
 "看看", "长什么样"
]
VIDEO_KEYWORDS = [
 "视频", "录像", "播放"
]
```

```
# 模型配置
EMBEDDING_MODEL = "tongyi-embedding-vision-plus"
LLM_MODEL = "qwen-flash"
```

关键参数配置

多模态统一向量空间



文本、图片、视频通过同一Embedding模型映射到统一向量空间，实现跨模态语义检索

检索策略：统一索引 + 后筛选

```
# 1. 统一检索 - 一次查询返回所有类型
results = search_with_details(query, index, metadata)

# 2. 意图检测 - 关键词判断是否需要媒体
want_image, want_video = detect_media_intent(query)

# 3. 分类筛选
# 文本: 无条件取Top-K
top_results = [r for r in results
        if r["type"] == "text"][:k]
```

```
# 图片: 仅当want_image且距离<3.0
if want_image:
    image_results = [r for r in results
        if r["type"] == "image"
        and r["distance"] < 3.0]
    matched_image = min(image_results,
                key=lambda x: x["distance"])
```

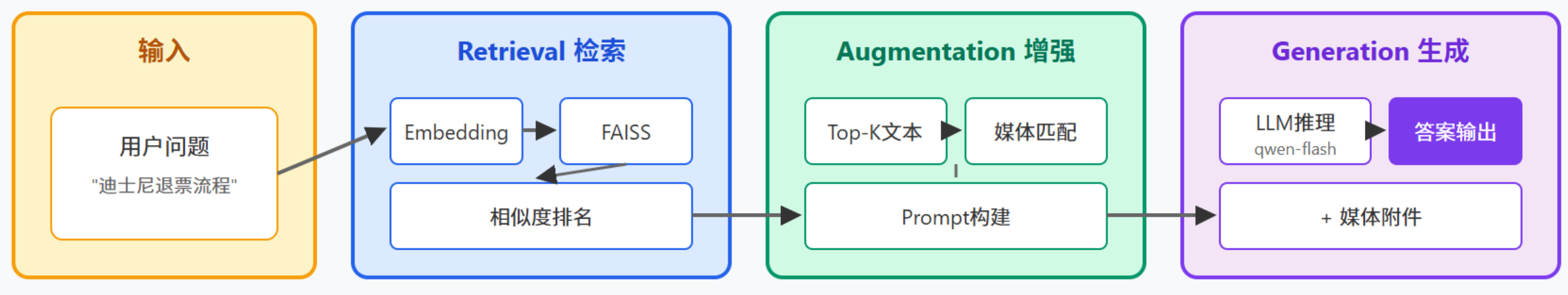
单索引架构：文本/图片/视频共享统一向量空间，使用同一个Embedding模型，索引维护简单

意图驱动：通过关键词检测用户意图，按需触发媒体检索，避免无关媒体干扰

文本优先：文本结果无条件进入Prompt，媒体作为附件补充，确保回答质量

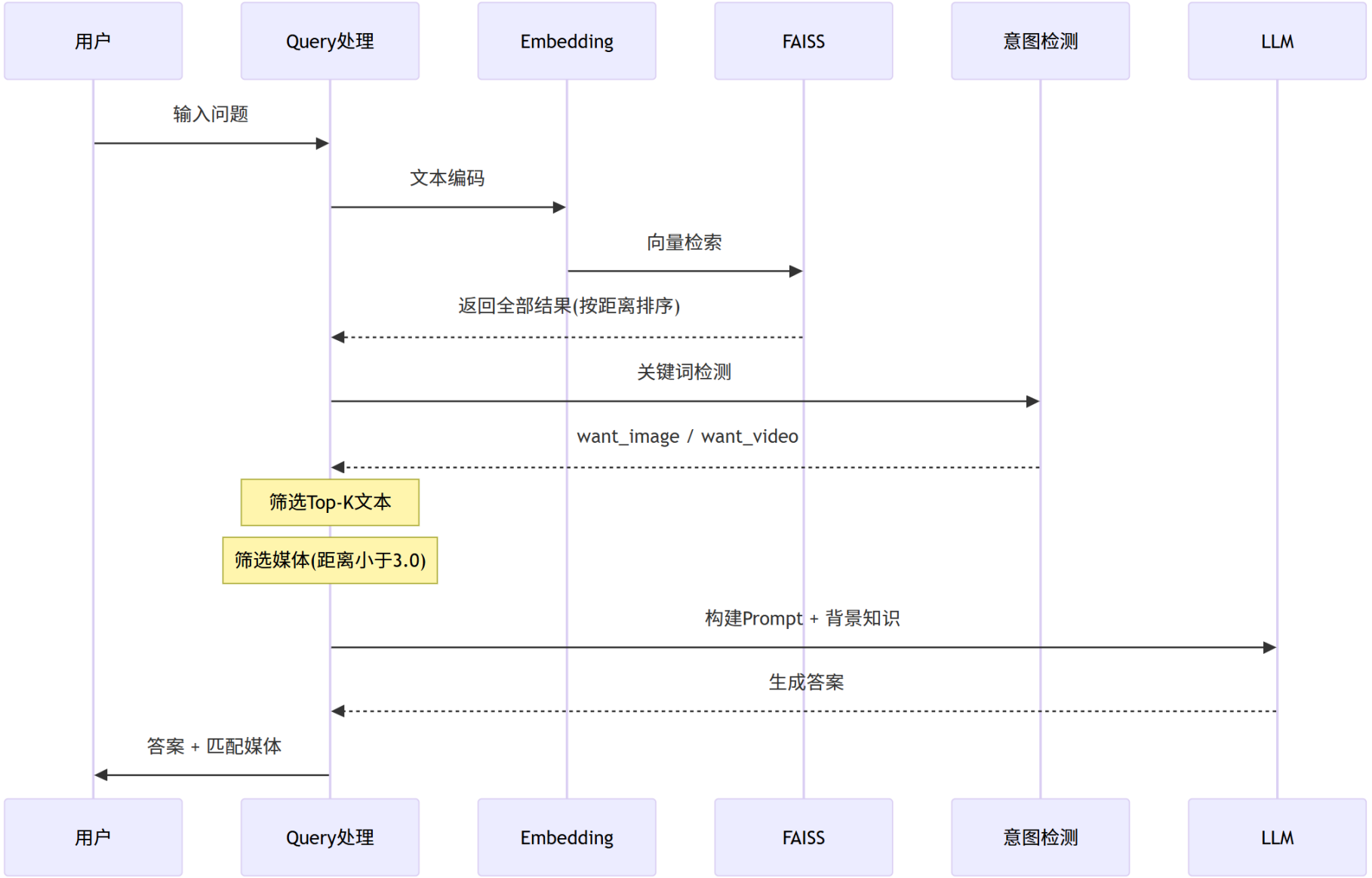
阈值过滤：媒体需满足距离<3.0才被采纳，防止低相关性媒体误匹配

Summary



RAG完整工作流

Summary（查询处理流程）



CASE：迪士尼RAG助手（索引创建）

--- 构建多模态知识库 ---
切分参数: chunk_size=500, overlap=50
处理文档: 1-上海迪士尼门票规则.docx
文档长度: 1342 字符, 切分为 3 个chunk
处理文档: 2-迪士尼老人票价规定.docx
文档长度: 879 字符, 切分为 2 个chunk
处理文档: 3-迪士尼乐园游玩攻略清单.docx
文档长度: 641 字符, 切分为 2 个chunk
处理文档: 4-上海迪士尼乐园酒店会员制度.docx
文档长度: 790 字符, 切分为 2 个chunk
处理图片...
- 1-聚在一起说奇妙.jpg
- 2-万圣节.jpeg
处理视频...
- 汽车剐蹭视频

向量维度: 1152
索引已保存: disney_index.faiss
元数据已保存: disney_metadata.json

完成! 文本:9, 图片:2, 视频:1

CASE：迪士尼RAG助手（用户问题1）

Query: 我想了解一下迪士尼门票的退款流程

==========

相似度排名 (越大越相似):

排名 ID 类型 相似度 距离 内容

---

1 2 [text] 0.5489 0.8220 章。储物柜分小型（60元/天）和大型（80元/天）两种，年卡用户享前两小时免费。童车租赁9...

2 8 [text] 0.5474 0.8267 留位置和烟花预留位置。礼宾服务会在上海迪士尼度假区官方app和微信公众号有售，最早提前7天...

3 0 [text] 0.5283 0.8927 上海迪士尼门票规则 上海迪士尼乐园门票分为一日票、两日票和特定日票三种类型。一日票可在 购买...

意图检测: 需要图片=False, 需要视频=False

选取Top-3文本构建Prompt:

- 章。储物柜分小型（60元/天）和大型（80元/天）两种，年卡用户享前两小时免费。童车租赁90元/天，... (相似度: 0.5489)

- 留位置和烟花预留位置。礼宾服务会在上海迪士尼度假区官方app和微信公众号有售，最早提前7天可购买，价... (相似度: 0.5474)

- 上海迪士尼门票规则

上海迪士尼乐园门票分为一日票、两日票和特定日票三种类型。一日票可在购买时选定日期... (相似度: 0.5283)

调用LLM生成答案...

最终答案:

您好！感谢您咨询上海迪士尼乐园门票的退款流程，以下是详细的说明，供您参考：

CASE：迪士尼RAG助手（用户问题1）

一、退票政策概览

1. **退改时间限制**：
   - 若您计划更改或取消门票，需在**入园前至少48小时**操作，方可享受免费修改或退票。
   - 距离入园日**不足48小时**（即2天内），将无法办理退改，系统将视为已使用。
2. **特殊情况处理**：
   - 如因突发疾病、意外事故等不可抗力情况导致无法入园，可提供**二级甲等及以上医院出具的诊断证明**，经核实后可申请退票。
   - 需在门票过期前30天内提交申请，并通过官方渠道提交相关材料。

3. **未使用门票延期**：
   - 门票若未使用且已过期，可在**过期后30天内**申请补差价延期。
   - 需支付原票价的**20%手续费**，具体金额根据实际票价计算。

二、如何办理退票/改期？

1. **操作方式**：
   - 登录【上海迪士尼度假区官方App】→【我的订单】→找到对应门票→选择“修改”或“退票”。
   - 或通过【上海迪士尼官网】或【官方微信公众号/小程序】进行操作。

2. **所需信息**：
   - 请确保购票时绑定的身份证件信息准确无误。
   - 退票成功后，款项将原路退回至您的支付账户（如支付宝、微信、银行卡等），通常需**3-7个工作日**到账。

...

CASE：迪士尼RAG助手（用户问题2）

Query: 最近万圣节的活动海报是什么
相似度排名 (越大越相似):

排名 ID 类型 相似度 距离 内容

---

1 8 [text] 0.4217 1.3712 留位置和烟花预留位置。礼宾服务会在上海迪士尼度假区官方app和微信公众号有售，最早提前7天...

2 2 [text] 0.4124 1.4250 章。储物柜分小型（60元/天）和大型（80元/天）两种，年卡用户享前两小时免费。童车租赁9...

3 1 [text] 0.4042 1.4742 后30天内可申请补差价延期，需支付原票价20%手续费。 年卡分为宝石卡、珍珠卡、翡翠卡三种...

4 6 [text] 0.4014 1.4910 ，注意保管好个人物品，尤其是贵重物品，如包包、相机等。16. 了解每个景点的适宜年龄范围，...

5 5 [text] 0.4013 1.4921 迪士尼乐园游玩攻略清单 1. 准备好迪士尼乐园的门票，可在官方销售点购买。2. 提前了解迪...

6 0 [text] 0.3971 1.5181 上海迪士尼门票规则 上海迪士尼乐园门票分为一日票、两日票和特定日票三种类型。一日票可在 购买...

7 7 [text] 0.3936 1.5406 上海迪士尼乐园酒店会员制度 首先迪士尼vip服务大致分为三种，一种是尊享卡（园区内十二个热...

8 10 [image] 0.3921 1.5501 [图片] 2-万圣节.jpeg... <-- 图片

意图检测: 需要图片=True, 需要视频=False

-> 匹配到图片: disney_knowledge_base\images\2-万圣节.jpeg (距离: 1.5501, 相似度: 0.3921)

选取Top-3文本构建Prompt:

- 留位置和烟花预留位置。礼宾服务会在上海迪士尼度假区官方app和微信公众号有售，最早提前7天可购买，价... (相似度: 0.4217)
- 章。储物柜分小型（60元/天）和大型（80元/天）两种，年卡用户享前两小时免费。童车租赁90元/天，... (相似度: 0.4124)
- 后30天内可申请补差价延期，需支付原票价20%手续费。 年卡分为宝石卡、珍珠卡、翡翠卡三种，价格从2... (相似度: 0.4042)

CASE：迪士尼RAG助手（用户问题2）

调用LLM生成答案...

最终答案:
您好！感谢您的提问～✨

关于上海迪士尼乐园近期万圣节活动的海报，目前官方已推出充满奇幻氛围的万圣节主题视觉设计，整体风格以“奇想之夜”为主题，融合了南瓜、幽灵、魔法与迪士尼经典角色的创意元素。海报中可以看到：

- 闪闪发光的南瓜灯点缀在城堡周围，营造出神秘又欢乐的节日气氛；
- 独特的“万圣节限定”角色形象，如米奇化身为小巫师、唐老鸭变成幽灵船长等；
- 重点突出万圣节夜场特别活动：**“奇想之夜”狂欢派对**，包括：
  - 惊喜不断的花车巡游（全新万圣节版）；
  - 城堡前的“点亮奇梦”光影秀升级为“幽灵奇境”特别版；
  - 限时开放的“惊魂密室”和“尖叫城堡”互动体验；
  - 迷你鬼屋与糖果大作战等亲子友好型趣味活动。

📌 **温馨提示**：
万圣节夜场票需**单独购买**，且仅限特定日期开放（通常为10月20日至11月5日左右），建议您通过【上海迪士尼度假区官方App】或【官方微信公众号】查看具体时间表，并提前购票哦！

如果您想第一时间获取最新海报及活动资讯，也可以关注我们的官方账号，我们会实时推送精彩内容～👻🎃

祝您玩得开心，愿这个万圣节充满惊喜与奇遇！💫

[相关图片]: disney_knowledge_base\images\2-万圣节.jpeg

CASE：迪士尼RAG助手（用户问题3）

Query: 我的汽车被剐蹭了，你能看到视频么？
相似度排名 (越大越相似):

排名 ID 类型 相似度 距离 内容

---

1 8 [text] 0.4120 1.4273 留位置和烟花预留位置。礼宾服务会在上海迪士尼度假区官方app和微信公众号有售，最早提前7天...

2 6 [text] 0.4087 1.4467 ，注意保管好个人物品，尤其是贵重物品，如包包、相机等。16. 了解每个景点的适宜年龄范围，...

3 2 [text] 0.4073 1.4552 章。储物柜分小型（60元/天）和大型（80元/天）两种，年卡用户享前两小时免费。童车租赁9...

4 11 [video] 0.3933 1.5424 [视频] 汽车剐蹭视频... <-- 视频

5 5 [text] 0.3903 1.5624 迪士尼乐园游玩攻略清单 1. 准备好迪士尼乐园的门票，可在官方销售点购买。2. 提前了解迪...

6 1 [text] 0.3843 1.6020 后30天内可申请补差价延期，需支付原票价20%手续费。 年卡分为宝石卡、珍珠卡、翡翠卡三种...

7 7 [text] 0.3812 1.6230 上海迪士尼乐园酒店会员制度 首先迪士尼vip服务大致分为三种，一种是尊享卡（园区内十二个热...

8 4 [text] 0.3766 1.6551 同者走快速通道。 票价调整机制方面，每年1月会根据物价指数进行微调，调幅通常不超过5%。重...

9 0 [text] 0.3681 1.7164 上海迪士尼门票规则 上海迪士尼乐园门票分为一日票、两日票和特定日票三种类型。一日票可在 购买...

10 3 [text] 0.3658 1.7340 迪士尼老人票价规定 迪士尼乐园针对老年游客提供票价优惠，具体规定如下。年龄要求方面，所 有年...

11 10 [image] 0.3477 1.8757 [图片] 2-万圣节.jpeg... <-- 图片

12 9 [image] 0.3381 1.9573 [图片] 1-聚在一起说奇妙.jpg... <-- 图片

意图检测: 需要图片=False, 需要视频=True

-> 匹配到视频: https://dataset-1255932437.cos.ap-nanjing.myqcloud.com/mp4/car.mp4 (距离: 1.5424, 相似度: 0.3933)

CASE：迪士尼RAG助手（用户问题3）

选取Top-3文本构建Prompt:

- 留位置和烟花预留位置。礼宾服务会在上海迪士尼度假区官方app和微信公众号有售，最早提前7天可购买，价... (相似度: 0.4120)

- ，注意保管好个人物品，尤其是贵重物品，如包包、相机等。16. 了解每个景点的适宜年龄范围，确保适合自... (相似度: 0.4087)

- 章。储物柜分小型（60元/天）和大型（80元/天）两种，年卡用户享前两小时免费。童车租赁90元/天，... (相似度: 0.4073)

调用LLM生成答案...

最终答案:

亲爱的游客，很抱歉听到您的汽车被剐蹭了。不过，作为迪士尼客服助手，我无法查看或提供任何监控视频。建议您立即联系乐园内的**游客服务中心**，工作人员会协助您记录事件并指引后续处理流程。

同时，请您：

- 保留现场照片或视频证据；
- 如有目击者，可留下联系方式；
- 若涉及保险理赔，请及时通知保险公司。

为了保障您的权益，也请在离园前尽快向游客服务中心报备相关情况。感谢您的理解与配合，祝您接下来的旅程愉快！✨

[相关视频]: https://dataset-1255932437.cos.ap-nanjing.myqcloud.com/mp4/car.mp4

切片策略

切片策略

Thinking：知识切片是RAG系统的核心环节，直接影响检索质量和回答准确性。都有哪些常用的切片策略？

方法1：固定长度切片

方法2：句子边界切片

方法3：LLM语义切片

方法4：层次切片

方法5：滑动窗口切片

切片策略

方法1：固定长度切片（有重叠）

- 按固定字符数切分文本，优先在句子边界进行切分，避免切断句子。
- 实现简单，处理速度快，长度统一，适合技术文档和规范文件。
- 适用场景：需要统一处理长度的场景，批量处理大量文档。

方法2：句子边界切片（无重叠）

- 基于自然语言处理，按句子、段落等语义单位进行切分。保持语义完整性，避免在句子中间断开，确保每个切片都是完整的语义单元。无重叠，按语义边界切分。
- 语义保持好，检索准确性高，但长度可能不均匀。
- 适用场景：适用于自然语言文本，需要保持语义完整性的场景。

切片策略

方法3：LLM语义切片

- 利用LLM的语义理解能力，在保持语义完整性的同时实现精确的长度控制。
- 语义理解能力强，分割点选择智能，但依赖GPU，成本较高。
- 适用场景：高质量要求的场景，复杂语义结构，有预算支持的项目。

方法4：层次切片

- 基于文档的层次结构（标题、章节、段落）进行切分。便于理解文档的逻辑关系。
- 保持文档结构，支持层次化查询，但依赖文档格式。
- 适用场景：结构化文档（手册、规范），多级标题的文档。

切片策略

方法5：滑动窗口切片

- 使用固定大小的窗口在文本上滑动，产生重叠的切片。通过重叠机制确保上下文连续性，减少信息丢失，提高检索召回率。
- 保持上下文连续性，减少信息丢失，但产生大量重叠内容。
- 适用场景：需要重叠信息的场景，长文档处理，需要保持上下文的场景。

切片策略对比

Thinking：针对某个知识库，不同的切片策略结果如何，优缺点是怎样的？

迪士尼乐园提供多种门票类型以满足不同游客需求。一日票是最基础的门票类型，可在购买时选定日期使用，价格根据季节浮动。两日票需要连续两天使用，总价比购买两天单日票优惠约9折。特定日票包含部分节庆活动时段，需注意门票标注的有效期限。

购票渠道以官方渠道为主，包括上海迪士尼官网、官方App、微信公众号及小程序。第三方平台如飞猪、携程等合作代理商也可购票，但需认准官方授权标识。所有电子票需绑定身份证件，港澳台居民可用通行证，外籍游客用护照，儿童票需提供出生证明或户口本复印件。

生日福利需在官方渠道登记，可获赠生日徽章和甜品券。半年内有效结婚证持有者可购买特别套票，含皇家宴会厅双人餐。军人优惠现役及退役军人凭证件享8折，需至少提前3天登记审批。

切片策略对比（固定长度切片）

方法1：固定长度切片

块 1 (292 字符):

迪士尼乐园提供多种门票类型以满足不同游客需求。一日票是最基础的门票类型，可在购买时选定日期使用，价格根据季节浮动。两日 票需要连续两天使用，总价比购买两天单日票优惠约9折。特定日票包含部分节庆活动时段，需注意门票标注的有效期限。

购票渠道以官方渠道为主，包括上海迪士尼官网、官方App、微信公众号及小程序。第三方平台如飞猪、携程等合作代理商也可购票，但需 认准官方授权标识。所有电子票需绑定身份证件，港澳台居民可用通行证，外籍游客用护照，儿童票需提供出生证明或户口本复印件。

生日福利需在官方渠道登记，可获赠生日徽章和甜品券。半年内有效结婚证持有者可购买特别套票，含皇家宴会厅双人餐。

块 2 (80 字符):

需在官方渠道登记，可获赠生日徽章和甜品券。半年内有效结婚证持有者可购买特别套票，含皇家宴会厅双人餐。军人优惠现役及退役 军人凭证件享8折，需至少提前3天登记审批。

1-固定长度切片.py

切片策略对比（句子边界切片）

方法2：句子边界切片

块 1 (287 字符):

迪士尼乐园提供多种门票类型以满足不同游客需求 一日票是最基础的门票类型，可在购买时选定日期使用，价格根据季节浮动 两日票 需要连续两天使用，总价比购买两天单日票优惠约9折 特定日票包含部分节庆活动时段，需注意门票标注的有效期限 购票渠道以官方渠道 为主，包括上海迪士尼官网、官方App、微信公众号及小程序 第三方平台如飞猪、携程等合作代理商也可购票，但需认准官方授权标识 所 有电子票需绑定身份证件，港澳台居民可用通行证，外籍游客用护照，儿童票需提供出生证明或户口本复印件 生日福利需在官方渠道登记，可获赠生日徽章和甜品券 半年内有效结婚证持有者可购买特别套票，含皇家宴会厅双人餐

块 2 (29 字符):

军人优惠现役及退役军人凭证件享8折，需至少提前3天登记审批

2-句子边界切片.py

切片策略对比（LLM语义切片）

方法3：LLM语义切片

```
prompt = f"""
请将以下文本按照语义完整性进行切片，每个切片不超过{max_chunk_size}字符。
要求：
1. 保持语义完整性
2. 在自然的分割点切分
3. 返回JSON格式的切片列表，格式如下：
{{
 "chunks": [
  "第一个切片内容",
  "第二个切片内容",
  ...
 ]
}}

文本内容：
{text}

请返回JSON格式的切片列表：
"""
```

切片策略对比（LLM语义切片）

LLM语义切片结果：

块 1 (57 字符): 迪士尼乐园提供多种门票类型以满足不同游客需求。一日票是最基础的门票类型，可在购买时选定日期使用，价格根据季节浮动。

块 2 (56 字符): 两日票需要连续两天使用，总价比购买两天单日票优惠约9折。特定日票包含部分节庆活动时段，需注意门票标注的有效期限。

块 3 (71 字符): 购票渠道以官方渠道为主，包括上海迪士尼官网、官方App、微信公众号及小程序。第三方平台如飞猪、携程等合作代理商也可购票，但需认准官方授权标识。

块 4 (50 字符): 所有电子票需绑定身份证件，港澳台居民可用通行证，外籍游客用护照，儿童票需提供出生证明或户口本复印件。

块 5 (54 字符): 生日福利需在官方渠道登记，可获赠生日徽章和甜品券。半年内有效结婚证持有者可购买特别套票，含皇家宴会厅双人餐。

块 6 (30 字符): 军人优惠现役及退役军人凭证件享8折，需至少提前3天登记审批。

3-LLM语义切片.py

切片策略对比（层次切片）

方法4：层次切片

块 1 (11 字符):

迪士尼乐园门票指南

块 2 (219 字符):

一、门票类型介绍

1. 基础门票类型

迪士尼乐园提供多种门票类型以满足不同游客需求。一日票是最基础的门票类型，可在购买时选定日期使用，价格根据季节浮动。两日票需要连续两天使用，总价比购买两天单日票优惠约9折。特定日票包含部分节庆活动时段，需注意门票标注的有效期限。

2. 特殊门票类型

年票适合经常游玩的游客，提供更多优惠和特权。VIP门票包含快速通道服务，可减少排队时间。团体票适用于10人以上团队，享受团体折 扣。

4-层次切片.py

切片策略对比（层次切片）

块 3 (214 字符):

二、购票渠道与流程

1. 官方购票渠道

购票渠道以官方渠道为主，包括上海迪士尼官网、官方App、微信公众号及小程序。这些渠道提供最可靠的服务和最新的票务信息。

2. 第三方平台

第三方平台如飞猪、携程等合作代理商也可购票，但需认准官方授权标识。建议优先选择官方渠道以确保购票安全。

3. 证件要求

所有电子票需绑定身份证件，港澳台居民可用通行证，外籍游客用护照，儿童票需提供出生证明或户口本复印件。

切片策略对比（层次切片）

块 4 (264 字符):

三、入园须知

1. 入园时间

乐园通常在上午8:00开园，晚上8:00闭园，具体时间可能因季节和特殊活动调整。建议提前30分钟到达园区。

2. 安全检查

入园前需要进行安全检查，禁止携带危险物品、玻璃制品等。建议轻装简行，提高入园效率。

3. 园区服务

园区内提供寄存服务、轮椅租赁、婴儿车租赁等服务，可在游客服务中心咨询详情。

生日福利需在官方渠道登记，可获赠生日徽章和甜品券。半年内有效结婚证持有者可购买特别套票，含皇家宴会厅双人餐。

军人优惠现役及退役军人凭证件享8折，需至少提前3天登记审批。

4-层次切片.py

切片策略对比（滑动窗口切片）

方法5：滑动窗口切片

块 1 (299 字符): 迪士尼乐园提供多种门票类型以满足不同游客需求。一日票是最基础的门票类型，可在购买时选定日期使用，价格根据季节浮动。两日 票需要连续两天使用，总价比购买两天单日票优惠约9折。特定日票包含部分节庆活动时段，需注意门票标注的有效期限。

购票渠道以官方渠道为主，包括上海迪士尼官网、官方App、微信公众号及小程序。第三方平台如飞猪、携程等合作代理商也可购票，但需 认准官方授权标识。所有电子票需绑定身份证件，港澳台居民可用通行证，外籍游客用护照，儿童票需提供出生证明或户口本复印件。

生日福利需在官方渠道登记，可获赠生日徽章和甜品券。半年内有效结婚证持有者可购买特别套票，含皇家宴会厅双人餐。军人优惠现役及

块 2 (173 字符): 小程序。第三方平台如飞猪、携程等合作代理商也可购票，但需认准官方授权标识。所有电子票需绑定身份证件，港澳台居民可用通行 证，外籍游客用护照，儿童票需提供出生证明或户口本复印件。

生日福利需在官方渠道登记，可获赠生日徽章和甜品券。半年内有效结婚证持有者可购买特别套票，含皇家宴会厅双人餐。军人优惠现役及退役军人凭证件享8折，需至少提前3天登记审批。

块 3 (23 字符): 退役军人凭证件享8折，需至少提前3天登记审批。

5-滑动窗口切片.py

切片策略对比

方法	核心思路	重叠	长度均匀	语义完整	实现成本	适用场景	不适用场景
固定长度切片	按字符数切，句子边界优化	有	高	中	低	通用场景、批量处理、对长度有要求	语义敏感的问答场景
句子边界切片	按句号分句，再合并	无	低	高	低	自然语言文本、问答系统	长句子多的文档
LLM 语义切片	LLM 理解内容后切分	无	中	最高	高	高质量要求、复杂语义结构	大规模文档、成本敏感
层次切片	按标题/章节切分	无	低	高	中	结构化文档（手册、规范、API 文档）	无标题的纯文本
滑动窗口切片	固定窗口滑动，大量重叠	大量	高	中	低	需要上下文连续、长文档召回	存储敏感、去重要求高

切片策略对比

Thinking：以下4种场景，你会选择哪种切分方式？

- 通用场景
- 技术文档
- 高质量要求
- 长文档召回

固定长度切片（简单可靠）

层次切片（保留结构）

LLM语义切片（效果最好）

滑动窗口切片（不漏信息）

打卡：迪士尼RAG助手

搭建完整的Disney RAG助手（原生RAG应用）

Step1. 数据层

- 文档处理：解析Word文档(.docx)，提取文本段落和表格（转为Markdown格式）
- 文本切分：按固定长度切分（500字符/chunk，50字符重叠）

Step2. 向量化层

- 统一Embedding：使用阿里云通义 tongyi-embedding-vision-plus 多模态模型
- FAISS单索引系统存储所有模态向量

Step3. 检索层

统一检索：文本query通过同一多模态模型编码，L2距离检索

关键词触发：检测特定关键词（如"海报"、"视频"）触发对应媒体类型过滤

Step4. 生成层

- Top-K文本召回构建上下文Prompt
- 自动附加匹配到的图片/视频链接


Thank You
Using data to solve problems