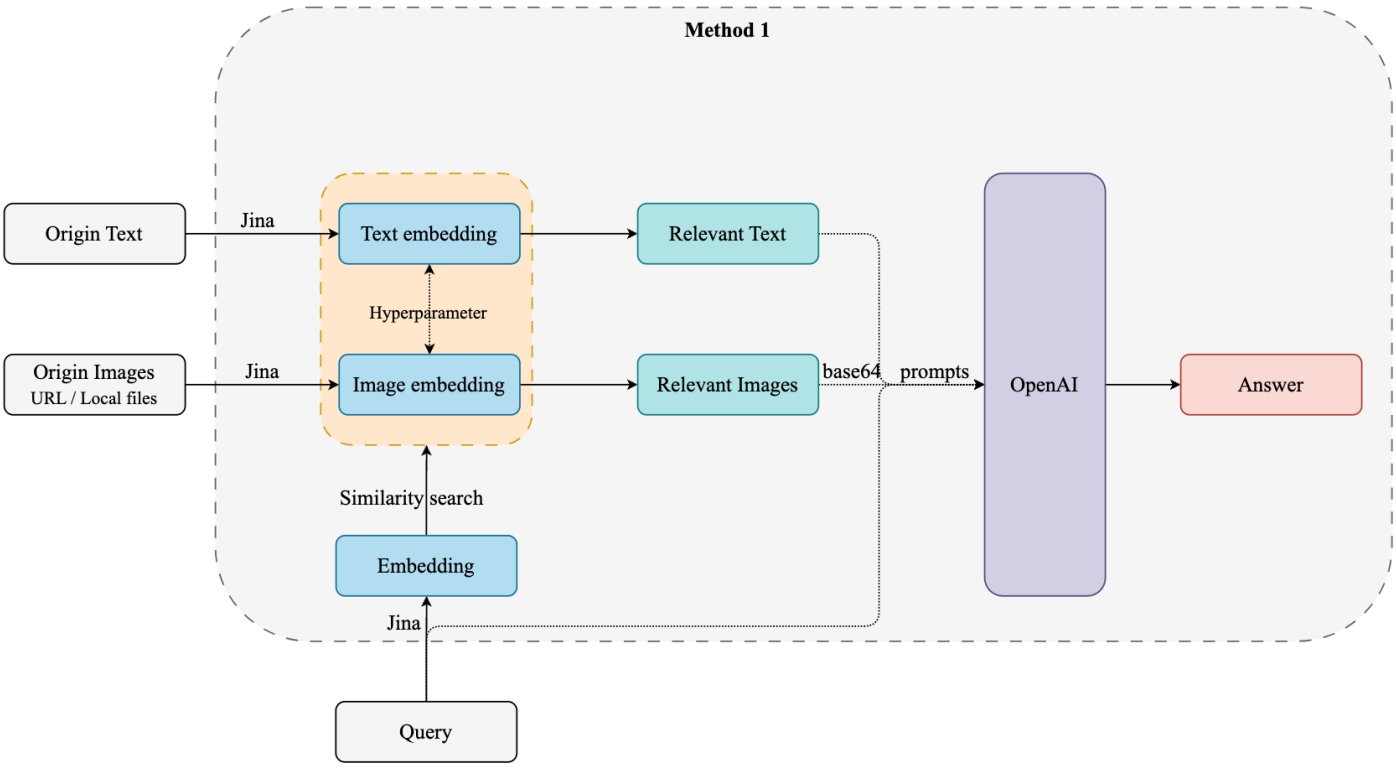
多模态检索

架构版本 1：Jina

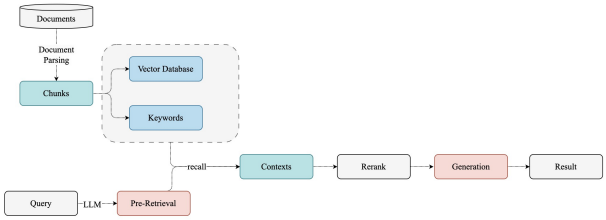
```
输入：
- query: str,  # 用户查询
- text: List[str], # 文本
- images: List[str], # 图像的文件路径

输出：answer

示例 demo：`test/multimodel_demo.ipynb`
```



说明：

原始数据：

- 三条文本texts = ['A pig', 'A red cat', "A red pig"]
- 两张图片：



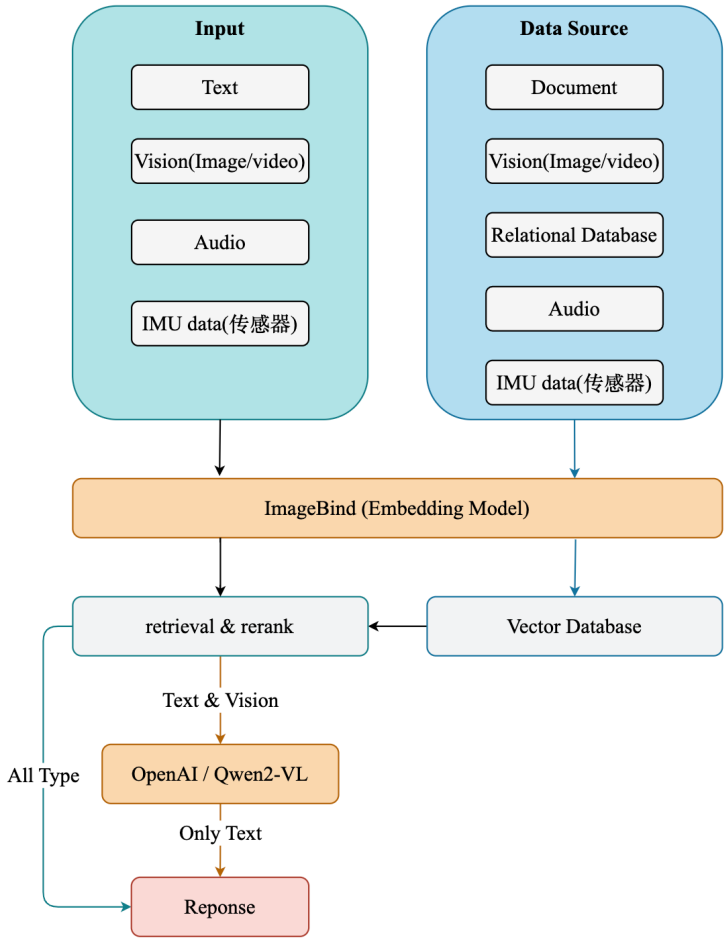
用户查询：小猪头发是什么颜色?

返回结果：小猪的头发是彩色的，包含红色、黄色和绿色。

用户查询：在 RAG 中recall 回 contexts 下一步应该做啥?

返回结果：在 RAG 中，recall 回 contexts 后，下一步应该进行 Rerank。

架构版本 2：Imagebind（lastest）



注意：

1. 在生成阶段，给推理模型的数据只能是 text / vision。无法用 audio 来做生成。
2. 可以解析 .gif，但是无法处理 .mp4.

局限性：

能用（文/图/音/传感数据）搜（文/图/音/传感数据）。

但生成阶段只能传给 LLM 文和图。Qwen 可以传 video，但是 `ImageBind` 编码不了 video，但能编码 .gif，但 Qwen 处理不了.gif。

架构版本 3: VLMs

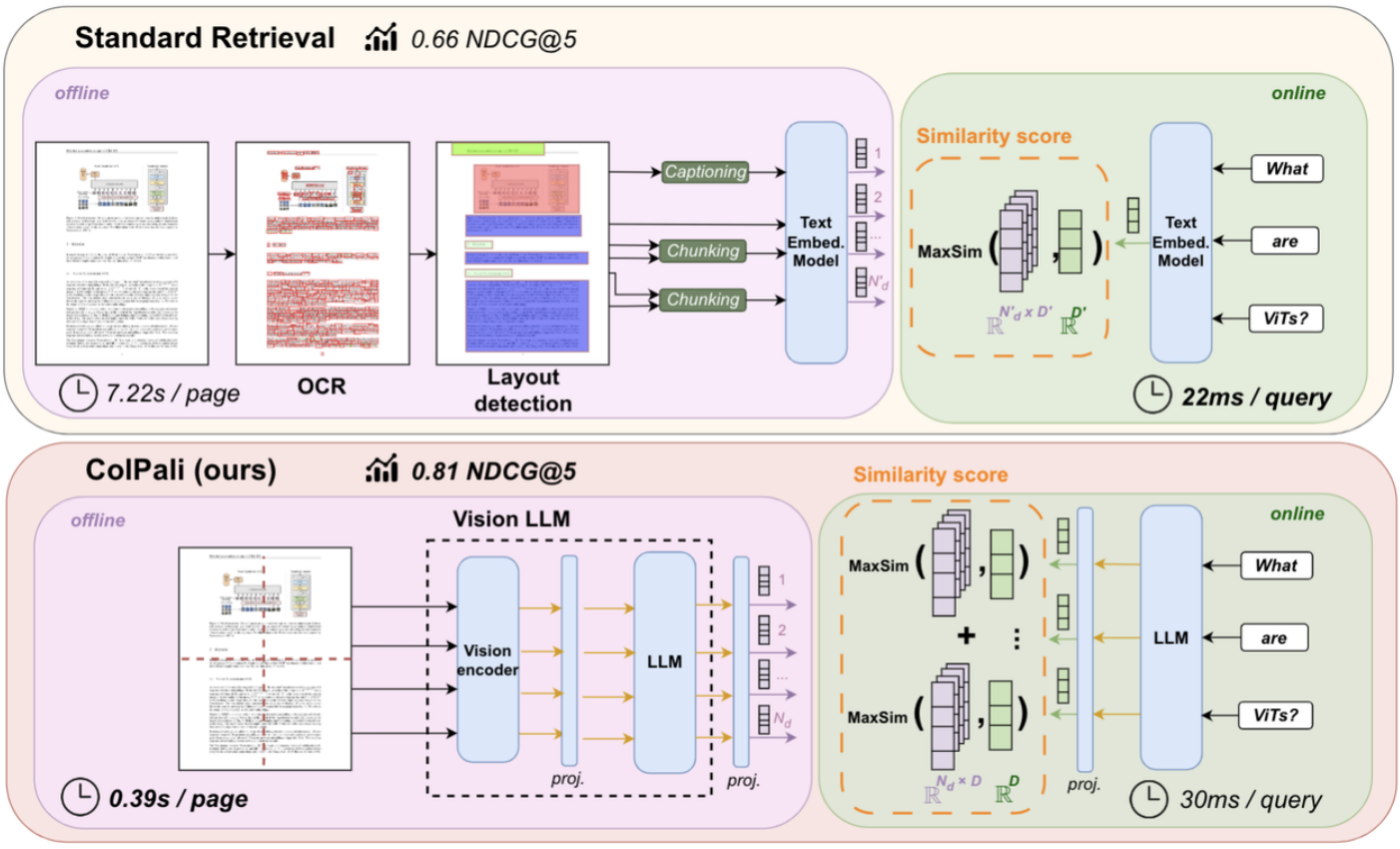
现有：处理 PDF，OCR + LLM

目前存在的问题：处理流程过于复杂。

使用OCR本身没有什么问题，但为什么要使用OCR，然后花时间去修正它，寻找重新编码布局信息的方法，重新格式化表格，检查一致性等等……而 VLMs 可以直接读取原始格式并解析所有内容。使用 VLMs 找出是否有什么因素会导致它在你的用例中失效（它们并不是完美的阅读器），这样你就可以更明智地分配资源来修复失败案例。

ColPali What:

利用 VLMs 的文档理解能力，仅通过文档页面的图像生成高质量上下文 embedding。结合 late interaction 匹配机制，ColPali大幅超越现代文档检索流程，同时速度更快且可端到端训练。



https://arxiv.org/pdf/2407.01449

https://huggingface.co/blog/manu/colpali

多模态 RAG：https://huggingface.co/collections/merve/multimodal-rag-66d97602e781122aae0a5139

Document AI：https://huggingface.co/collections/merve/awesome-document-ai-65ef1cdc2e97ef9cc85c898e

如果你想要从结构相同的文档中获得简短、结构化和简洁的答案，并且有标记的数据，建议对像 Donut 或 LayoutLM 系列或 UDOP 这样的模型进行微调。

OCR 模型：

https://github.com/VikParuchuri/surya?tab=readme-ov-file#benchmarks

VLM：如果要定位和回顾生成答案的引用时，好像无法提供坐标来定位。

x.com

https://x.com/jonathanrlarkin/status/1831640797843767438

Reference：

Compositional Chain-of-Thought Prompting for Large Multimodal Models

github.com

文搜图处理

1. 根据图像和文本任务生成场景图。
2. 接下来，通过使用图像、场景图、问题和答案提取提示来提取答案。
3. 最后，生成的场景图是对图像进行紧凑语言表示。