视频截屏识别需求 全部的截屏画面收集：[视频截屏画面收集](https://doc.weixin.qq.com/sheet/e3_ADMA-gZjABocg6fQ4nfTLWV07hMb3?scode=AIAAlgcWAAkeL4WEG5ADMA-gZjABo&tab=0g465s)

# 需求

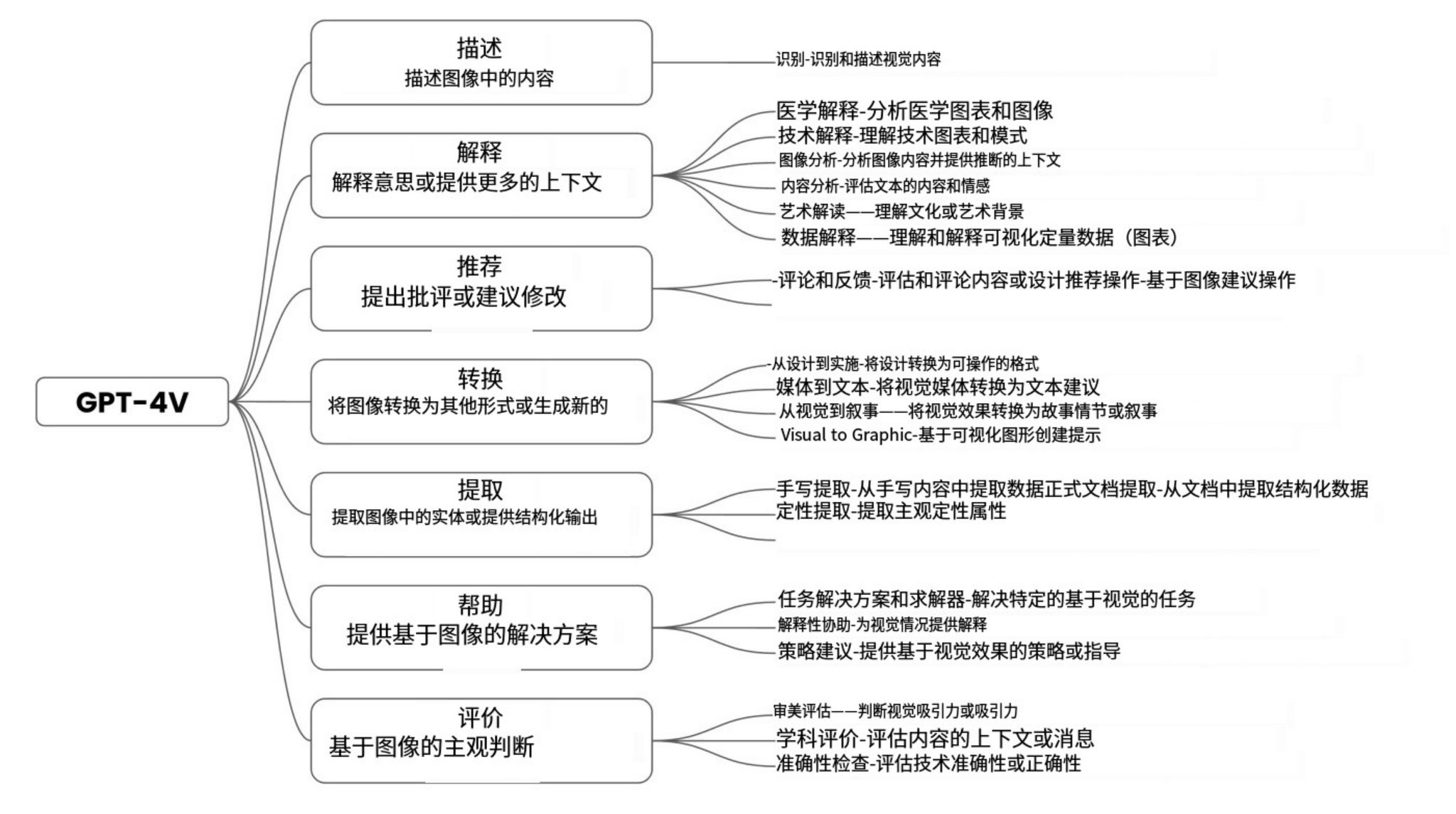
主要满足 图片描述 和 图片信息提取 两种场景；

描述的用户需求是方便日后搜索；

提取的场景是把图片上的信息提取为文字笔记，一般常见的是：将图片上的文本内容存为半结构化的笔记，或者准确提取 URL，公式等，代替用户手工敲字；

图片描述为我们的主要需求，提取为次要需求。

下图是非官方总结的GPT-4V的七种常见图片处理需求。



## 设计

1. 具体产品设计时，可能会采用 描述+提取 一体化的设计，但要看提取的效果如何再定最终设计。首先描述图片，同时把OCR提取的内容，经过整理排版也给到用户。
2. 描述图片，对准用户搜索查找场景；采用描述模版，规范对图片的描述，避免大模型过多的解读带来的错误信息；具体格式：标题 + 摘要 + 涉及 + 链接、公式。

|  |  |  |
| --- | --- | --- |
| 模版项 | 实现目标 | 内容规范参考 （具体示例看后面的表格) |
| 标题 | 概括图片主题 | 识别内容中 文字大小，以最大字为解读素材，如果判断最大字体是含义充分的（如含义不充分，则加上下一级大小的文字），则直接使用；如果字体大小统一，则对全部可解读文字做极简摘要，提示大模型用做卡片标题；字数不超过30字，一般15字左右比较常见。 |
| 摘要 | 简要概括图片内容，不做过多的解读和内容逻辑上的描述，避免出错。 | 针对所有识别的文字，做一句话概括，字数不超过60字（字数待定，60是目前已测试后得出的初步判断，需详细测试后明确） |
| 涉及 | 增加更多图片内的实体关键词，便于搜索，也方便用户复制使用。 | 将文字中的概念、人名、地名、时间等概念或实体通过罗列的方式描述出来；如：  涉及：   * 人名：罗素、图灵 * 概念：数学原理、组合决策 * 组织：MIT、哈佛 |
| 链接&公式 | 正确提取 | 对图片上的链接和公式，需要单独提取出来，展示在特定区域，保证正确性。 |

1. 提取截屏时刻的字幕给到大模型，加强对图片理解。需要判断截屏这一帧的停留时长，合理截取截屏前+截屏停留中的字幕，这块需要讨论；具体效果如何，需要测试。

## 后续产品设计上的一些细节

1. 保存截屏时，需要将以下信息一起保存，以扩大搜索范围：视频名称，提取的原文等，具体还有哪些，产品设计时确定。
2. 自动翻译：如用户目标语言与所看视频语言不一样，自动将提取内容翻译成目标语言，并同时显示两种语言的文字。
3. 用户保存截屏时，可以手动选择字幕，跟截屏一起保存为知识卡片。

## 截屏内容处理示例

|  |  |  |
| --- | --- | --- |
| **参考类型** | **示意图** | **描述参考（手工编写的，供参考，结合实际大模型能实现的程度）** |
| 仅非结构化文字 | descript | 标题：基础模型不是“助手”  摘要：基础模型不回答问题，而是完整的互联网文档。可以操纵它们来执行快速工程任务。  涉及：   * 概念：模型、提示工程 |
| 仅复杂不易解图形 | descript | 这张内容少，用来测试，看看效果 标题：编程语言  摘要：  涉及： |
| 复杂不可理解图形+零散文字 | descript | 标题：tokens和ULMFiT三步法  摘要：关于tokens和ULMFiT三步法的内容  涉及：   * 概念：tokens、预训练、微调、分级微调、ULMFiT |
| 复杂不可理解图形+零散文字 | descript | 标题：符号处理、逻辑计算从1900年到1950年的历史  摘要：符号处理、逻辑计算从1900年到1950年的历史  涉及：   * 人名：罗素、阿尔弗雷德·诺斯·怀特海德、图灵、阿隆佐·丘奇、埃米尔·L·波斯特 * 概念：数学原理、Entscheidungsproblem、形式化简、组合决策问题、符号操作 |
| 对比 | descript | 标题：GPT-3 和 LLAMA 语言模型之间的对比  摘要：GPT3（2020）与LLAMA（2023）在大小、架构、优化超参数等方面的对比，涉及GPT3和LLAMA各对应一个数据表格  涉及：   * 概念：GPT3、LLAMA、超参数 |
| 结构化（表格） | descript | 标题：神经网络的历史可以追溯到几十年前，为什么会重新兴起呢?  摘要：神经网络的复兴，归因于大数据、改进的硬件和软件以及更容易构建这些网络的代码等因素  涉及：   * 概念：神经网络、大数据、 * 组织：MIT |
| 排名 | descript | 标题：热门应用每月用户突破1亿花了多长时间  摘要：关于热门应用每月用户突破1亿花了多长时间的排名  涉及：   * 概念：chatgpt、tiktok、Instagram、Spotify、Uber、telegram、MAUs |
| 演变 | descript | 标题：GPT 和其他语言模型的演变  摘要：概述了从 2018 年到 2023 年包括 GPT 在内的各种语言模型的演变。它重点介绍了开源和闭源模型。  涉及：   * 概念：GPT、xx、xx（其他模型名称）   网址：[github.com/Mooler0410/LLMsPracticalGuide](http://github.com/Mooler0410/LLMsPracticalGuide) |
| 流程 | descript | 标题：GPT Assistant 训练流程  摘要：关于GPT Assistant 训练流程的概述，详细介绍了预训练、监督微调、奖励建模和强化学习等阶段。  涉及：   * 概念：GPT、预训练、微调、算法、大模型 |
| 公式 | descript | 标题：感知器：前向传播  摘要：来自麻省理工学院（MIT)的关于深度学习的讲座或教程，特别关注感知器前向传播的概念  涉及：   * 概念：非线性输出 * 组织：MIT |
| 手写 | descript |  |

# 现有问题改进

|  |  |  |
| --- | --- | --- |
| 类型 | 问题描述 | 图 |
| 稳定性 | 我们对同一张截屏的多次处理差别大，trickle是稳定的 | descript  descript  descript  descript  descript |
| 翻译质量 | token是否不翻译更好 | descript |
| Chain of thought的固定翻译：思维链 | descript |
| 过于泛化 | 我们有时事实提取的不好，泛化过了 | descript  descript |
| 排查问题 | 在[《trickle测试》](https://doc.weixin.qq.com/sheet/e3_ADMA-gZjABocg6fQ4nfTLWV07hMb3?scode=AIAAlgcWAAkyCd4AbwADMA-gZjABo&tab=gke6jq)表格里，需要排查的图片 | descript |
|  |  |  |

全部的截屏画面收集：[视频截屏画面收集](https://doc.weixin.qq.com/sheet/e3_ADMA-gZjABocg6fQ4nfTLWV07hMb3?scode=AIAAlgcWAAkeL4WEG5ADMA-gZjABo&tab=0g465s)