作品名称

LLLLL；余天男；徐炀；冉辉；邱刚

（不要出现任何涉及学校名称等内容）

摘要

本项目设计并开发了一款基于RK3588芯片的AI智能声景人机交互导盲项链，旨在提升现有导盲设备在复杂路况中的识别准确率及人机交互能力。通过对现有导盲设备市场的调研发现，目前大多数产品功能单一，难以适应复杂多变的户外环境，且缺乏高效、自然的人机交互机制，不能满足视障人群在实际出行中的多样化需求。

针对上述问题，本项目提出了一种双模式工作机制：对话模式与导盲模式。在对话模式下，用户可通过集成的语音交互模块实现与设备的自然语言交流，完成如模式切换、状态查询、指令输入等操作，提高使用的便捷性与智能化水平；在导盲模式下，设备通过搭载的摄像头采集周围环境图像，并结合基于Qwen-vl多模态大模型，对红绿灯、行人、车辆等关键障碍物进行精准识别。同时，系统会精简的利用喇叭告诉用户需要注意的路况，引导用户安全通过复杂环境。

核心硬件平台选用高性能AI芯片RK3588，利用其强大的算力支持高精度图像识别与实时语音处理，实现边缘计算能力在嵌入式导盲设备中的集成应用。整个系统以佩戴式项链为载体，兼顾功能性与可穿戴性，适合视障用户长时间日常使用。

本项目的设计有效融合了人工智能、语音识别、人机交互与嵌入式系统等关键技术，提升了导盲设备的智能化水平与用户体验，具有良好的社会意义与推广应用价值。

第一部分　作品概述

1.1 功能与特性

1.1.1 双模式工作机制

对话模式：集成语音识别与自然语言处理模型 whisper，实现用户与设备之间的语音交互。用户可通过语音进行自由对话，提升操作便捷性与交互体验。

导盲模式：基于 Qwen-vl 多模态大模型，定时用摄像头拍照，然后通过语音描述路况，引导用户安全通行。

1.1.2 高性能计算平台

系统核心硬件采用 RK3588 AI 芯片，具备强大的计算能力和图像处理性能，支持高速神经网络推理，实现设备本地实时处理，降低延迟并提升稳定性。

1.1.3 多目标识别与威胁提示

摄像头定时拍照显示路况，经过 Qwen-vl 多模态大模型分析后可识别多类目标，并结合距离、移动方向与位置关系判断其对用户的潜在威胁程度，进行语音分级播报，帮助用户快速理解环境状况，增强出行安全。

1.1.4 语音反馈系统

内置高灵敏度麦克风与扬声器，语音播报清晰，提示内容具有优先级。结合环境音识别调节播报音量，提高在嘈杂环境下的识别率和可听性。

1.2 应用领域

应用于视障人群的日常出行场景，特别是在交通路口、人流密集区域、复杂室外环境中提供主动预警和方向引导。替代传统白手杖、导盲犬等方式，实现更高精度的**智能识别**与**语音提醒**。同时可作为智慧城市中的辅助可穿戴终端，接入城市交通系统（如红绿灯联网）、导航系统，实现环境感知与人机互动结合。在无障碍出行系统中作为重要组成部分，提高城市对弱势群体的服务水平。 项目基于 **RK3588 芯片**和 Qwen-vl 多模态大模型的结合，展示了**边缘侧 AI 推理**在实际产品中的落地能力，适用于移动终端视觉识别；本地语音交互控制；嵌入式 AI 设备原型开发。另外可作为公益项目推广对象，被用于康复中心、残障辅助

机构的导盲辅具试点。在政府或慈善组织采购中具有潜力，助力“科技助残”工程。

1.3 主要技术特点

1.3.1 高性能边缘计算平台

搭载**瑞芯微 RK3588 AI** 芯片，内置 NPU，具备强大算力，支持本地深度学习推理。同时支持多路高清视频处理，满足实时图像识别与语音处理双任务的需求。

1.3.2 语音交互系统

结合语音识别（Whisper）与自然语言处理，能够与用户进行语音互动，提升操作便捷性和交互体验。双向语音系统支持用户与设备进行自然对话，提升可操作性和用户体验。

1.3.3 多目标识别与威胁提示

利用 Qwen-vl 模型对拍摄的图像进行多目标识别，结合物体距离、移动方向等信息判断潜在威胁，并通过语音分级播报，让用户快速理解周围环境，提升出行安全

1.3.4 双模式智能工作机制

对话模式：用于语音交互、设置和状态查询； 导盲模式：专注于环境感知与导航辅助，自动进入任务模式，处理图像并语音播报。

1.3.5 本地处理，无需联网

所有识别与处理任务均在设备端完成，无需依赖云端，不依赖网络连接，确保低延迟和隐私安全。

1.4 主要性能指标

项目名称	指标参数描述
处理器芯片	瑞芯微 RK3588，8 核 CPU + 内置 6 TOPS NPU
操作系统	Ubuntu 系统
目标识别算法	Qwen-vl 多模态大模型
处理时间	本地部署，相比云端模型交互快 2-3 倍

项目名称	指标参数描述
工作模式	对话模式、导盲模式双模式切换
威胁评估功能	Qwen-vl 多模态模型自动评估威胁等级
摄像头规格	1080P 高清摄像头，广角 ≥ 120°
语音交互模块	支持本地语音识别与播报，降噪麦克风 + 高保真扬声器
供电方式	可充电锂电池，支持 5V 供电，续航 ≥ 6 小时（视实际功耗）
通信接口	USB-C / UART / GPIO（根据原型可选）
设备形态	可穿戴项链式结构，轻量化设计，整机重量 ≤ 200g
适用环境	室内外通用，支持-10℃~50℃ 工作温度，防尘防滴溅（IP4X）

1.5 主要创新点

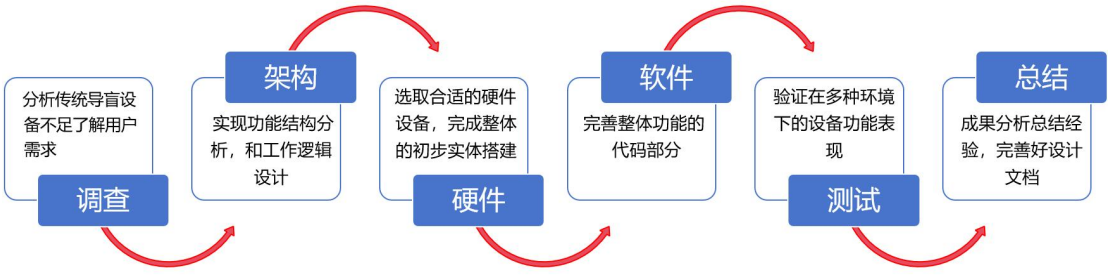
(1) 基于边缘计算的本地智能视觉识别系统，实现实时图像处理与多目标检测。

(2) 引入 Qwen-vl 多模态模型，识别很多种物体的情况下，自主评估威胁等级，能完整播报整个路况中盲人需要知道的信息，比如红绿灯，台阶，坑洼等等

(3) 设备支持对话模式和导盲模式两种模式，实现真正意义上的“听觉+视觉 AI 辅助融合”。

(4) 完全离线运行，隐私安全友好。所有图像识别与语音处理功能均在本地完成，无需联网，提升用户信任感和使用安全性。

1.6 设计流程



阶段一（调查）：分析传统视障人群的出行痛点和现有导盲设备交互的不足，了解用户需求。

阶段二（架构）：确定总体功能架构，实现语音，视觉双模式的工作逻辑设计。

阶段三（硬件）：选择摄像头、麦克风、扬声器等外设，进行结构建模与样机

3D 打印。

阶段四（软件）：完善总体功能代码，同时完成 AI 模型的部署和语音交互的功能。

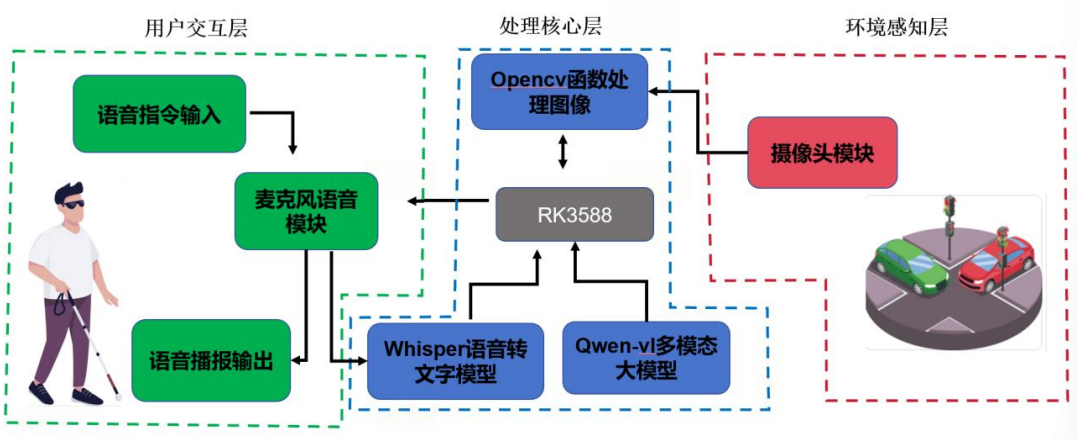
阶段五（测试）：联合调试图像识别、语音输入输出模块，验证不同场景下识别稳定性与响应速度

阶段六（总结）：编写设计文档和报告书。

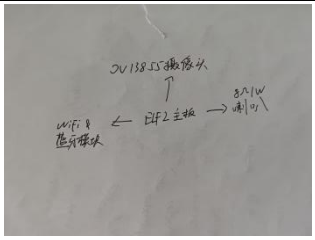
第二部分　系统组成及功能说明

2.1 整体介绍

这个系统是为视障人士设计的智能辅助工具，结合语音交互和环境感知技术，旨在帮助用户更好地感知和理解周围的环境。系统通过摄像头实时捕捉图像，利用 Qwen-vl 模型进行目标检测，识别整个路况，找出潜在的障碍物或危险，评估环境中的威胁，并根据这些信息做出响应。同时，用户通过语音输入与系统互动，发出指令或询问，系统通过语音播报提供反馈，帮助用户了解周围情况或进行相应操作。同时接入的 Qwen-vl 模型还可以实现轻量化的 AI 对话。整个系统的核心处理单元依托 RK3588 处理器，确保数据处理和响应的高效性，为用户提供精准和及时的辅助，提升视障人士的生活质量和独立性。



2.2 硬件系统介绍



2.2.1 硬件整体介绍：；

2.2.3 电路各模块介绍：

OV13855 摄像头

OV13855 是 OmniVision Technologies 生产的一种 13 兆像素 CMOS 图像传感器，专为主流智能手机设计。它支持 4K2K 视频（45 fps）、1080p 高清（60 fps）等功能，接口为 MIPI，功耗低（活跃模式约 233mW）

8 欧姆 1W 喇叭

这是一种 8 欧姆阻抗、1 瓦功率的音频扬声器，适合小型电子项目。常见于 Arduino 或 Raspberry Pi 的音频输出，频率响应通常为 600 Hz 至 10 kHz。；

2.3 软件系统介绍

2.3.1 软件整体介绍（含 PC 端或云端，结合关键图片）；

2.3.2 软件各模块介绍

主控制模块：该模块由 main()函数实现，负责通过 OpenCV 打开摄像头采集图像，保存至本地，并利用文件通信机制向 AI 推理服务发送图像路径指令，随后循环等待响应锁文件出现以确认 AI 服务完成图像分析，读取分析结果文件内容，并调用文本转语音模块播报反馈。关键输入包括摄像头设备索引及文件通信路径，关键输出为生成的图像文件路径（写入命令文件）和接收到的文本描述结果（读取响应文件）。

AI 推理服务模块：以 C++中的 main()为入口，该模块实现视觉编码模型及多模态大语言模型（LLM）的初始化，读取命令文件中图像路径，对图像进行预处理（函数 expand2square()）、缩放，调用 run_imgenc()进行图像编码，构造完整的文本视觉联合输入后调用 rkllm_run()执行推理，通过回调函数 callback()异步接收生成文本，推理结束后将文本写入响应文件，并创建锁文件通知主控模块。关键输入为图像文件路径及模型配置参数，关键输出为完整的文本推理结果和响

应状态文件。

语音录制与识别模块：其核心函数是 run_recording_and_transcribing()，通过调用录音脚本生成音频文件，然后调用 Whisper 模型进行语音转文本识别，返回识别结果文本；模块中 speak()函数实现对识别文本的语音合成播报，save_latest_text()函数将识别文本保存到指定文件，供系统其他模块使用。关键输入是麦克风采集的语音信号和录音配置，关键输出为对应的文本命令及语音确认播放。

文本转语音模块：由 speak(text)函数实现，通过调用外部语音合成工具 espeak-ng 将文本转换成 wav 音频文件，并通过 aplay 命令播放生成的音频文件，播放完成后删除临时文件。该模块的关键输入是待播报的文本字符串，输出经由系统音频设备播放的合成语音。

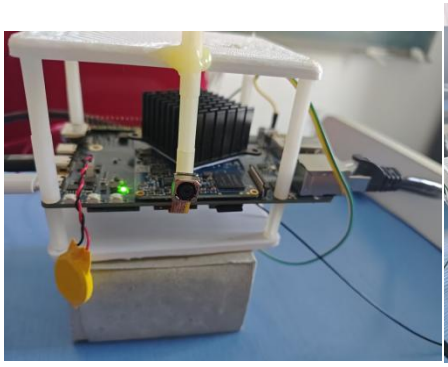
文件通信模块：贯穿于主控与 AI 推理服务的多个函数中，利用预定义的命令文件（存储图片路径或指令文本）、响应文件（存储 AI 生成结果）、响应锁文件（标识响应完成）以及就绪信号文件（标识服务已启动），该模块实现进程间的异步通信和状态同步。读写操作发生于主控程序 main()与 AI 服务模块 main()函数中，关键输入为指令内容，关键输出为响应文本及同步信号文件；

第三部分　完成情况及性能参数

阐述最终实现的成果

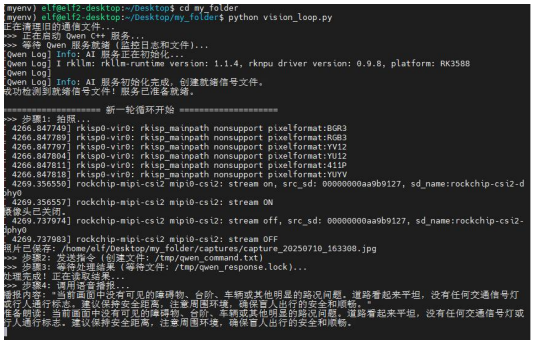
摄像头定时拍照，传入给 Qwen-vl 多模态大模型进行路况整体分析，并输出盲人可能需要的信息，在用喇叭输出出来

3.1 整体介绍




3.2 工程成果（分硬件实物、软件界面等设计结果）

3.2.1 软件成果



；

第四部分 总结

4.1 可扩展之处

在感知层面，可增加红外、激光雷达及视觉辅助模块，提升环境感知的精度，范围与实时性，以适应复杂出行场景。交互上，通过语义识别完成复杂任务指令。系统通信拓展至 4G/5G 网络，实现与手机 App、云端服务器的实时数据同步，具备远程监控、亲属位置共享、后台路线推荐等功能。终端形态适配手杖、手环、背包等硬件，提升佩戴与使用舒适度；输出方式兼顾视障用户感知差异，提供可调节震动节奏、多声道音效等选择。这些人性化设计将推动系统持续演进为更智能、个性化且贴合用户真实需求的综合辅助系统。

4.2 心得体会

这次大学生团队参加嵌入式大赛的经历，让我们深刻体会到创新、合作和技术融合的重要性。我们设计并开发了一款基于 RK3588 芯片的 AI 智能声景人机交互导盲项链，旨在提升导盲设备在复杂路况中的识别准确率及人机交互能力。回顾整个研发过程，每个环节都充满了挑战与成长。

在项目初期，我们进行了市场调研，发现市面上的导盲设备大多功能单一，无法应对复杂的户外环境。因此，我们决定设计一个结合 AI 技术的智能导盲项链，能够实时感知和适应环境变化，更好地帮助视障人士出行。

硬件方面，我们选用了 RK3588 芯片，因其强大的算力与图像处理能力，支持高效边缘计算，能够在实时语音处理和图像识别上提供优异的表现。利用这一平台，系统通过摄像头采集环境图像，并结合 Qwen-vl 模型进行目标检测，精准识别红绿灯、行人、车辆等障碍物，可根据路况变化提供语音预警，帮助用户安全通过复杂环境。

在软件开发中，团队成员分工明确，负责不同模块的研发。我主要负责语音交互模块的设计与优化，确保用户能够通过自然语言与设备进行高效沟通，完成指令输入、模式切换等操作，提升设备的智能化水平。其他成员则专注于视觉识别与系统优化，通过多次调试和优化，确保各模块兼容与系统稳定运行。

项目开发过程中，我们遇到了不少跨学科的挑战，特别是在系统集成与调试阶段。图像识别、语音识别与硬件优化等技术的结合，对我们提出了高要求。但通过团队的紧密合作，我们一步步攻克了这些难题，确保了项目的顺利进行。

这次比赛让我深刻认识到团队合作的重要性。每个团队成员都发挥了自己的优势，大家互相帮助、共同解决问题，确保项目成功实施。此外，嵌入式技术的应用让我体会到跨学科知识的魅力，实践中我们不仅应用了硬件和软件的结合，还涉及了图像处理、语音识别等多个领域的知识。

尽管取得了一些成绩，但我们知道这只是开始，如何将这一技术推广到更多视障人士的日常生活中，仍是我们未来的目标。通过这次大赛，我们不仅提高了技术能力，也积累了宝贵的实践经验。这为我们未来的职业生涯提供了重要的启示，也将继续激励我们在未来的工作中，不断创新、跨学科合作，解决实际问题。

第五部分 参考文献

调查报告：

Whisper 模型的文献调研

Whisper 是由 OpenAI 开发的一种语音识别模型，旨在通过大规模弱监督数据实现高鲁棒性和多语言转录能力。经搜索和分析，Whisper 的主要参考文献为 Alec Radford 等人于 2023 年发表的论文《Robust Speech Recognition via Large-Scale Weak Supervision》。这篇论文在《第 40 届国际机器学习会议》（ICML 2023）上发表，并可在 arXiv 上获取（链接：https://arxiv.org/abs/2212.04356）。

论文详细描述了 Whisper 模型的训练过程，指出其使用 680,000 小时的多语言和多任务监督数据进行训练，表现出在零样本转移设置下的良好泛化能力，与人类在准确性和鲁棒性上的表现接近。论文还提到模型和推理代码已开源，为进一步研究提供了基础。

此外，搜索结果还包括多个相关资源，如 OpenAI 官网的介绍（https://openai.com/index/whisper/，发布于 2022 年 9 月 20 日）、GitHub 仓库（https://github.com/openai/whisper，更新至 2024 年 9 月 30 日）、Hugging Face 上的模型详情（https://huggingface.co/openai/whisper-large-v3）以及 Wikipedia 页面（https://en.wikipedia.org/wiki/Whisper_%28speech_recognition_system%29，更新至 2023 年 8 月 12 日）。这些资源进一步验证了论文的权威性，并提供了模型的实际应用和性能评估。

例如，Hugging Face 页面提到 Whisper large-v3 在超过 500 万小时标记数据上训练，显示出在多种语言上的 10%至 20%的错误减少，相比之前的 large-v2 版本。

AWS 博客

（https://aws.amazon.com/blogs/machine-learning/whisper-models-for-automatic-spe

ech-recognition-now-available-in-amazon-sagemaker-jumpstart/，发布于 2023 年 10 月 10 日）也讨论了 Whisper 在 Amazon SageMaker JumpStart 中的应用，提供了性能指标如字错误率（WER）的比较。

Qwen-VL 模型的文献调研

Qwen-VL 是由 Alibaba Cloud 的 Qwen 团队开发的多模态大语言模型，专注于视觉和语言的联合理解。经搜索，Qwen-VL 的主要参考文献为 Qwen 团队于 2023 年 8 月 24 日上传至 arXiv 的论文《Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond》（链接：https://arxiv.org/abs/2308.12966）。

论文介绍了 Qwen-VL 系列模型的设计，包括视觉接收器、输入输出接口、三阶段训练管道和多语言多模态清洗语料库。模型不仅支持传统的图像描述和问答，还实现了图像-标题-框元组的对齐，增强了定位和文本阅读能力。在多个视觉中心基准测试（如图像字幕、问答、视觉定位）中，Qwen-VL 和 Qwen-VL-Chat 模型在零样本和少样本设置下创下新纪录。

搜索结果还包括 Qwen-VL 的 GitHub 仓库（https://github.com/QwenLM/Qwen-VL，发布于 2023 年 8 月 21 日），描述其为“Qwen 最强大的大型视觉语言模型”，以及 Qwen 官网的介绍（https://qwenlm.github.io/blog/qwen-vl/，发布于 2024 年 1 月 25 日）。这些资源进一步确认了论文的地位，并提供了模型的更新版本如 Qwen2-VL 和 Qwen2.5-VL 的详细信息。

例如，GitHub 页面提到 Qwen-VL-Max 在中文问答和文本理解任务上优于 OpenAI 的 GPT-4V 和 Google 的 Gemini。Encord 的博客（https://encord.com/blog/qwen-vl-large-scale-vision-language-models/，发布于 2024 年 2 月 29 日）讨论了 Qwen-VL 在 OpenVLM 排行榜上的领先地位，涵盖 38 个视觉语言模型的 13 个多模态任务。