

Citywalk场景中的 VideoAgent设计与实践

演讲人：宋阳

OPPO / 算法专家

AiCon
全球人工智能开发与应用大会

目录

01

AI实景对话总体框架

VideoAgent应用

02

Citywalk交互设计

AI探店助手

03

Citywalk实战挑战

意图指代与记忆

04

Citywalk未来演进

提前感知和主动推荐

05

VideoAgent行业观察

热点agent介绍

06

极客邦科技 2026 年会议规划

促进软件开发及相关领域知识与创新的传播



参会咨询



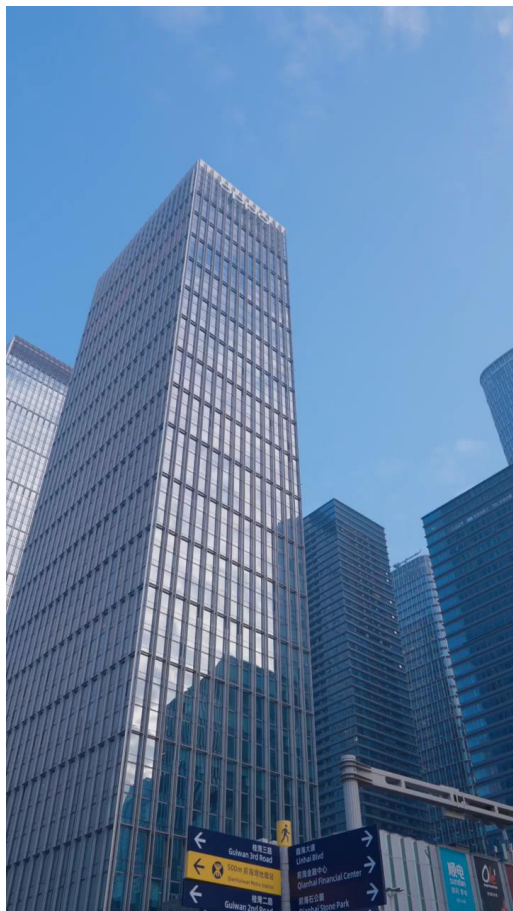
查看会议



01 AI实景对话总体框架

VideoAgent应用

AI实景对话



AI实景对话是OPPO ColorOS16推出的重点系统级AI功能，它让手机秒变“行走的万事通”。

用户只需将手机镜头对准目标，无论是餐厅菜单、商品吊牌还是展览说明，轻点屏幕即可触发智能识别，支持连续对话与随时打断，即使在商场嘈杂环境中，也能通过声纹识别精准响应指令。

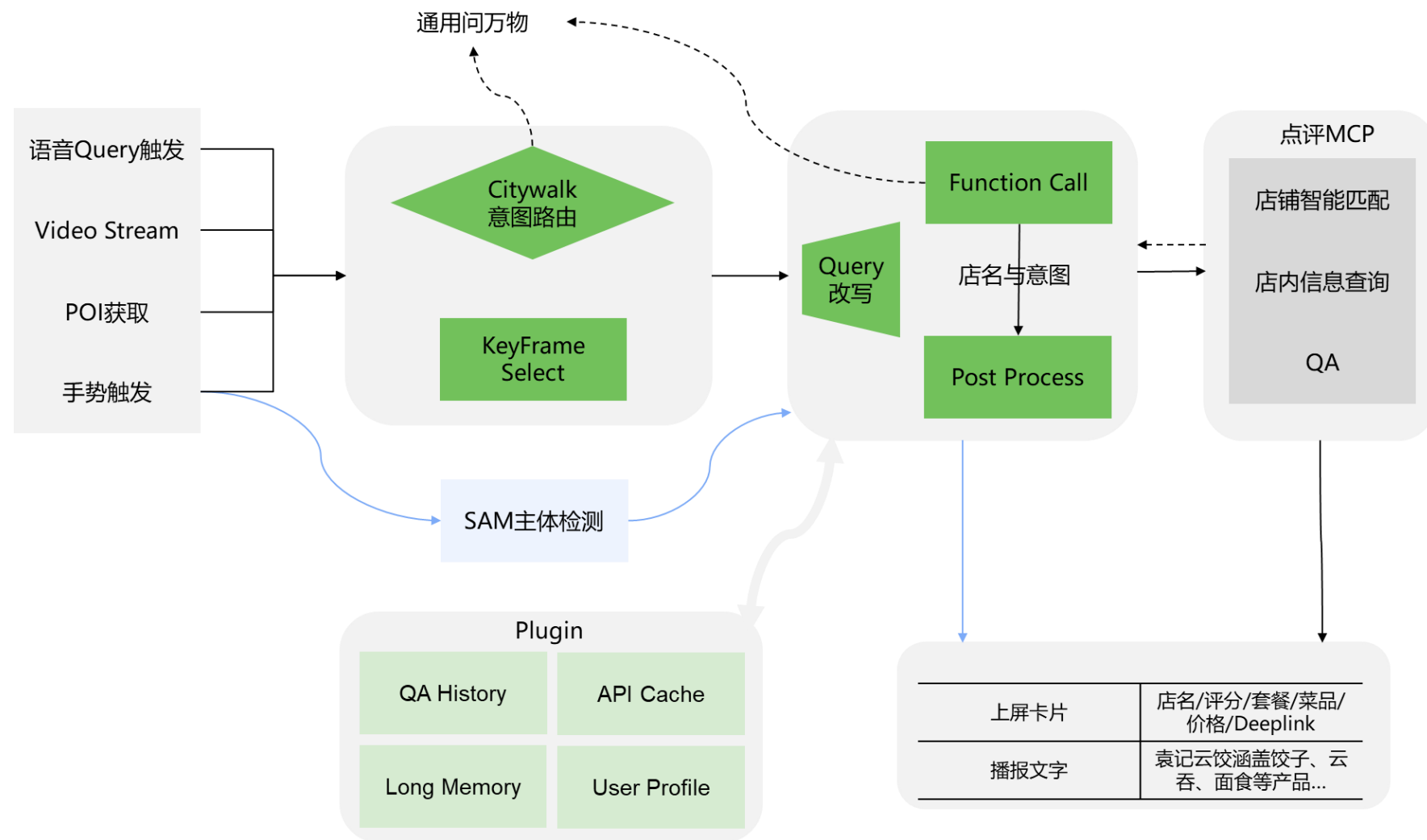
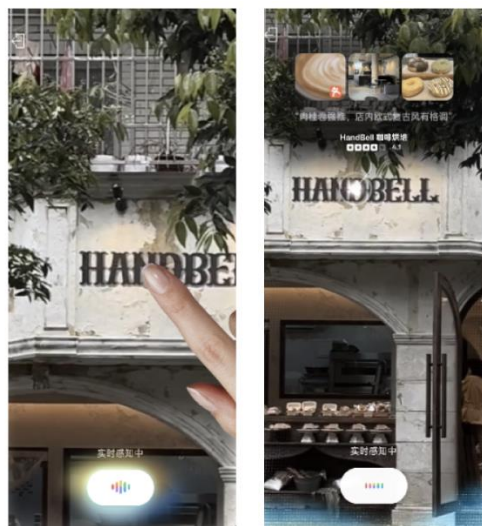
技术链条



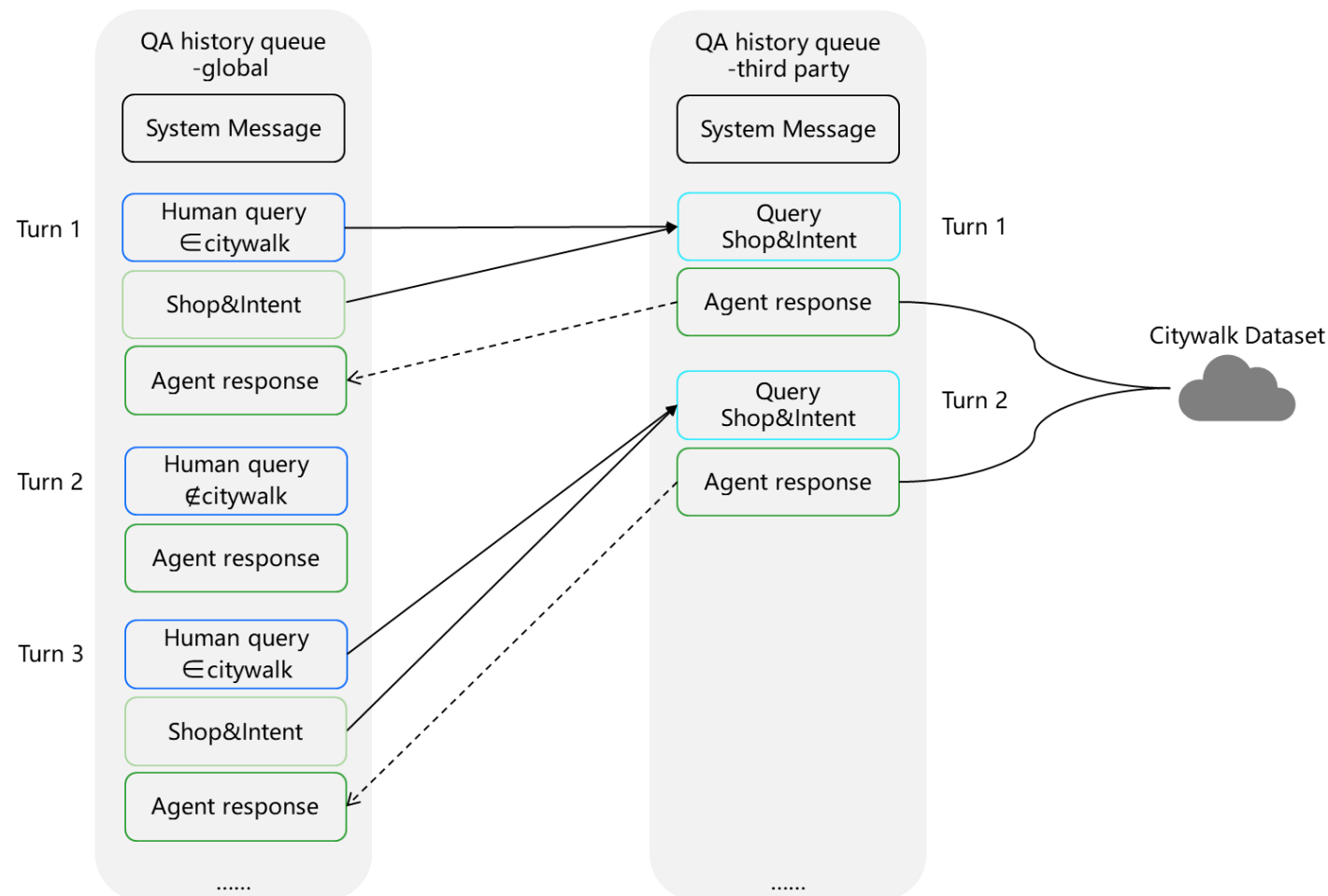
02 Citywalk交互设计

AI探店助手

Citywalk交互

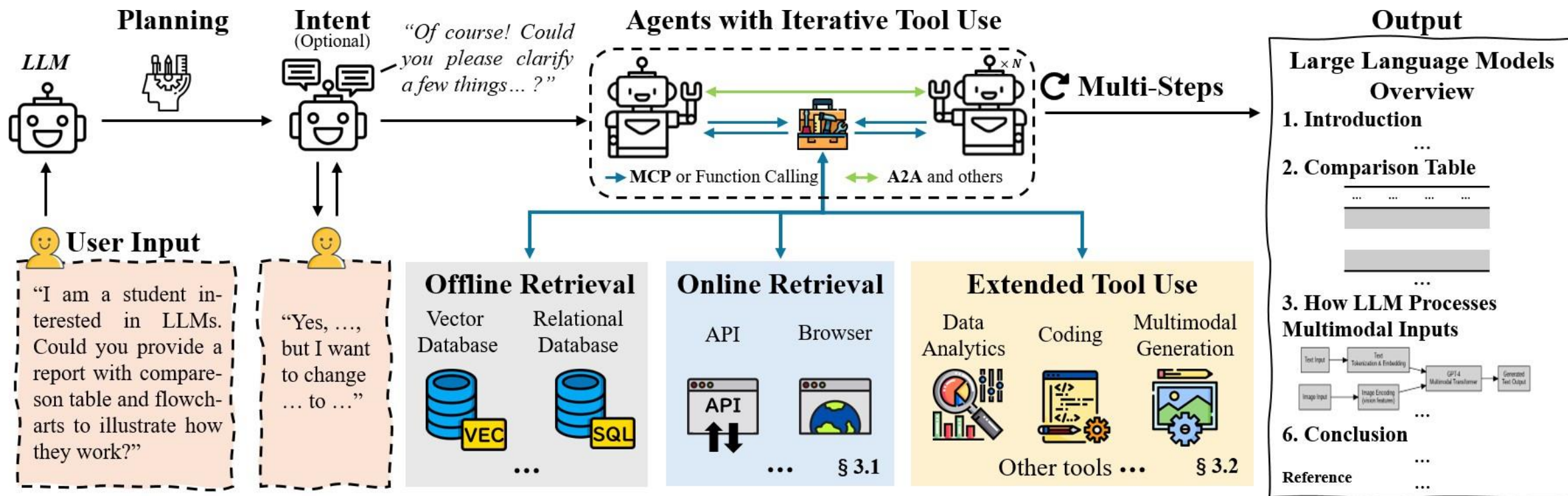


Citywalk交互



对于citywalk意图的轮次，同步传输query和店名意图，异步取回回复结果，并记录对应的userid、recordid和turnnum。

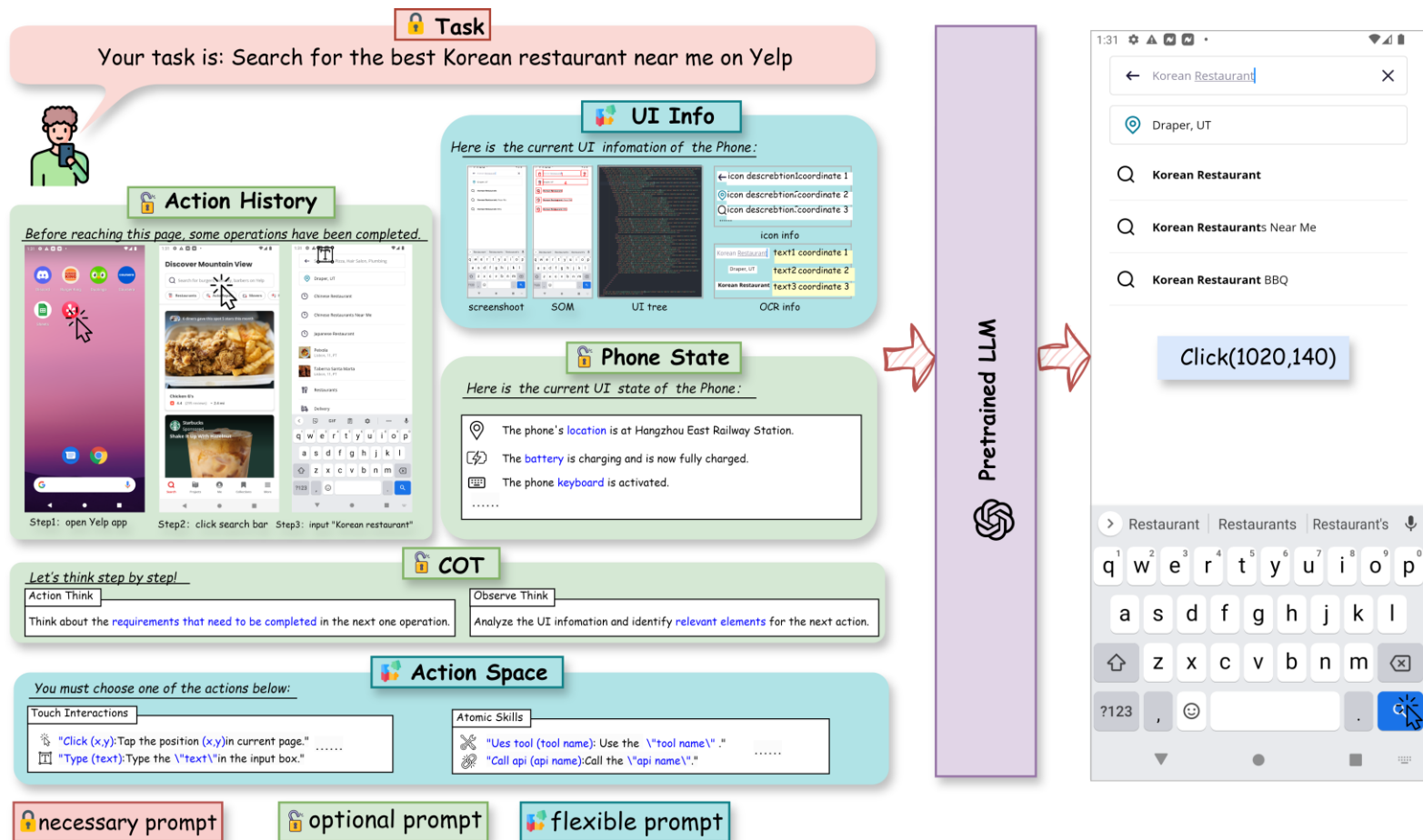
对比Deep ReSearch



集成动态推理、自适应规划、多轮信息检索与工具调用能力，最终生成结构化分析报告。

—Deep Research Agents: A Systematic Examination And Roadmap.

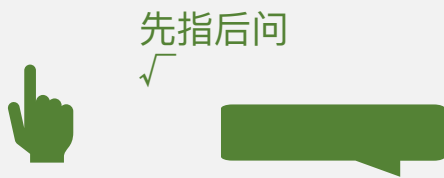
对比GUI Agent



将用户意图映射到UI操作。GUI Agent实时感知UI信息和手机状态，基于已有知识和记忆进行规划、推理和反思，最后执行操作。

—LLM-Powered GUI Agents in Phone Automation: Surveying Progress and Prospects.

指问结合



先指后问
✓



仅指 ✓

边指边问
✓



先问后指 ×

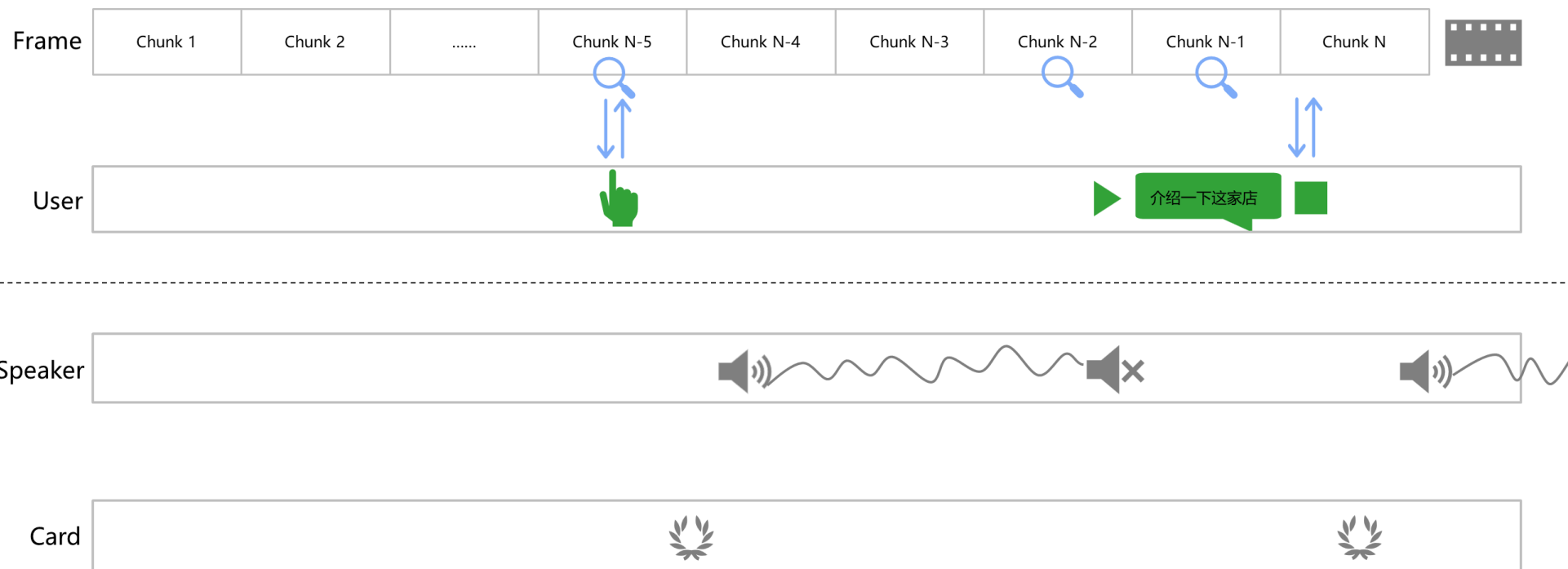






仅问 ✓



仅问的情况识别牌匾面积占比大的店铺，手指点击可以锁定目标店铺更精准。
先问后指的情况，会先针对query回复，在语音播报时手指屏幕，会打断处理。

关键帧选择



-  KeyFrame Select
-  Frame synchronization
-  VAD Start
-  VAD End

点击情况：点击帧就是关键帧，按fps均匀采样后获取最近时刻的帧；

语音query情况：VAD起始时刻附近的帧是关键帧，关键帧通常发生在附近2个Chunk内；

实时打断：TTS播报过程中，如果用户提问或点击，系统将打断原播报并进入下一轮交互。

03 Citywalk实战挑战

意图指代与记忆

意图粒度

店铺通用

介绍以下这家店/包厢有没有5个人的
这是什么/这家店电话多少/可以停车吗

消费价格

这家店贵吗/4个人吃饭需要多少钱
这家店我吃的起吗/什么时候有折扣



菜品情况

他们家主要吃什么/排名第一个是什么菜
主推菜是什么/有没有适合小朋友的菜

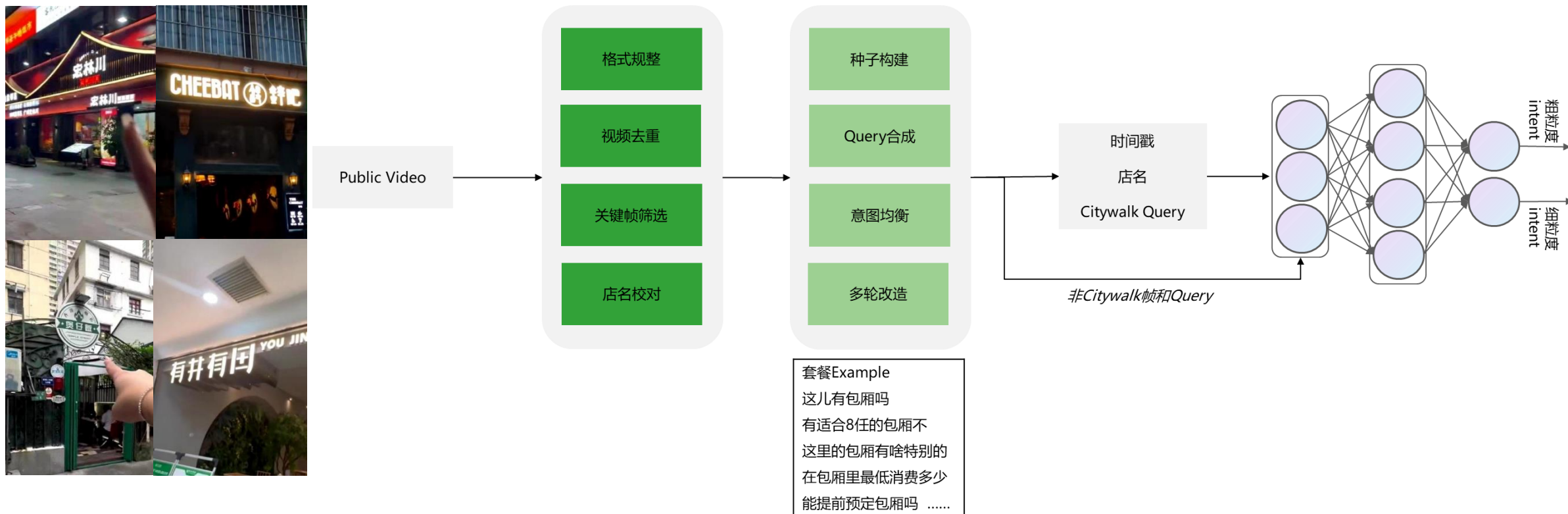
评论分数

这家店评分咋样/这家店口味如何
这家店环境如何/服务态度好不好

粗粒度意图关注的是当前交互是否属于citywalk，是二分类问题；

细粒度意图关注的是用户query的问题类型，是多分类问题，并最终映射到多种样式的上屏卡片。

意图识别

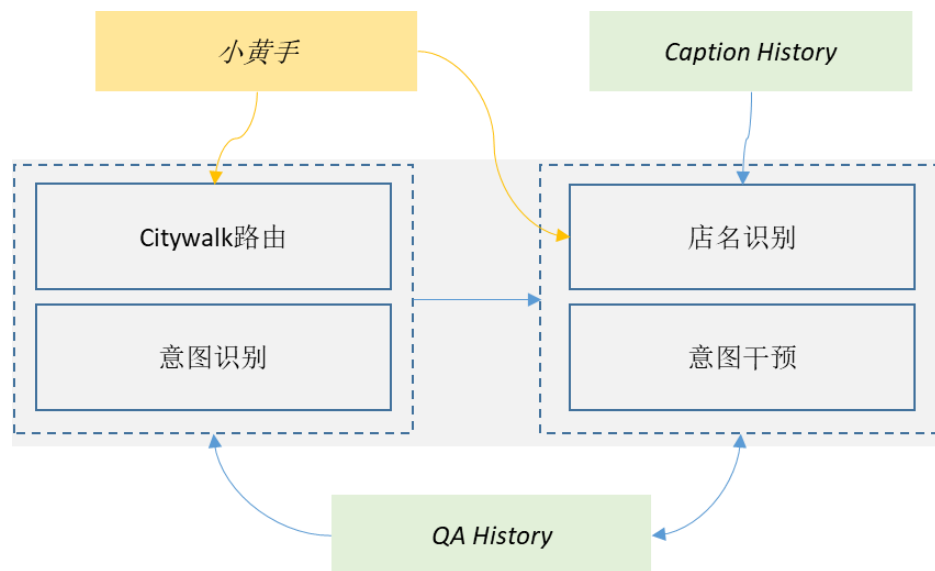


TrainSet-image: 覆盖餐饮、服务业、娱乐类，负例是通用场景画面；

TrainSet-query: 基于强模型合成citywalk query，负例是citywalk无关问题；

TestSet: 后训练InternVL4B模型取得95%的准确率，优于闭源大模型few-shot效果（83%）

记忆回溯



TimeStamp	Shop \subseteq KeyFrame	QA History	Caption History
Chunk N	×	-	-
Chunk N-1	√	√	√
Chunk N-2	×	√	√
Chunk N-3	√	√	√
...



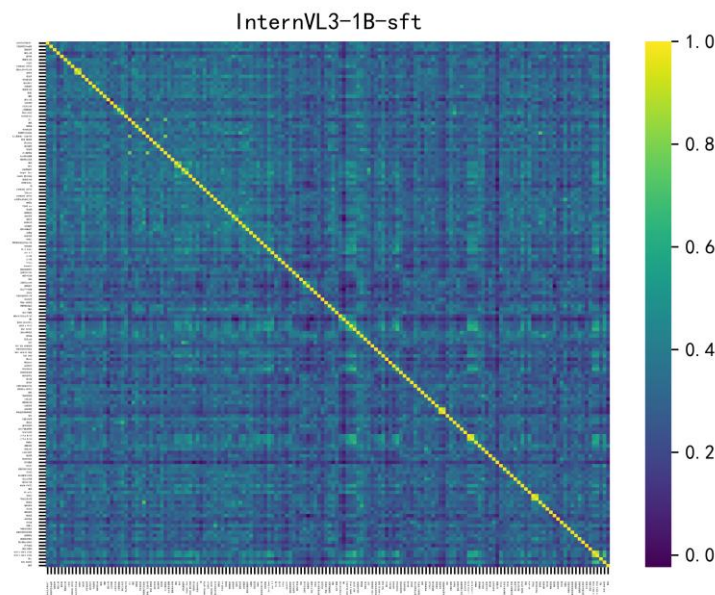
当关键帧未出现店名时，根据时间轴倒序回溯历史QA和历史Caption，避免答非所问。
即使用户镜头没有一直对着门匾拍摄，助手也可以根据记忆锚定店铺并给出准确答案。

记忆回溯



对于点击+单轮query情况，通常回溯到历史qa的上一轮即可，点击帧无需重复识别；
对于单轮query情况，通常需要回溯历史caption，找到距离当前时刻最近的店铺；
对于多轮query情况，需要对历史qa和历史caption回溯判断。

后处理



黔夺夺贵州酸汤火锅 (远大广场店)

4.0

火锅 ¥70/人

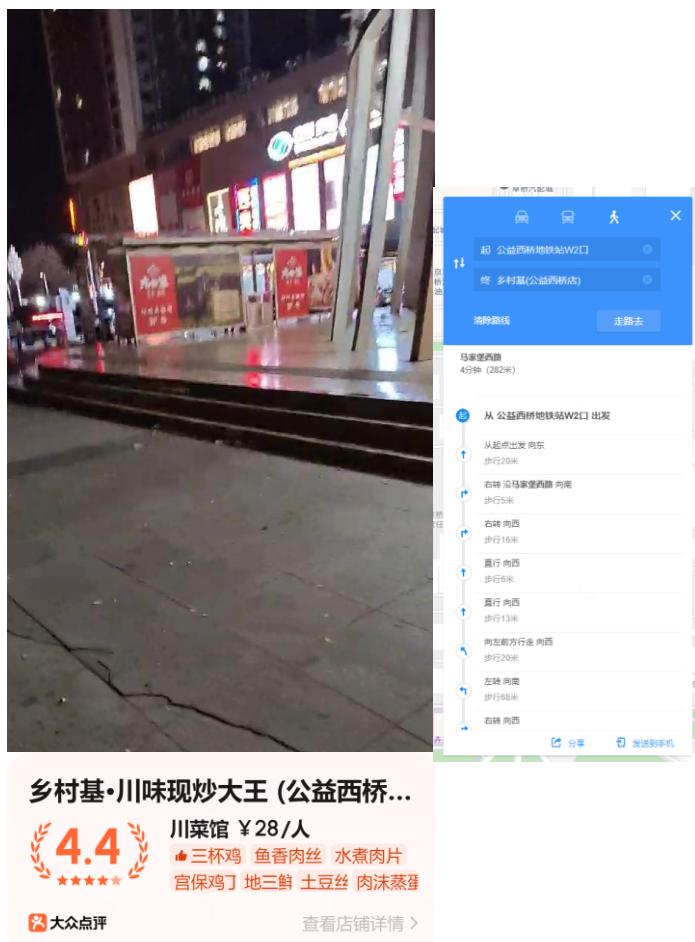
贵州酸汤 山野木姜子牛肉
脆哨洋芋粿 贵阳糯米饭 牛肉拼盘

大众点评

查看店铺详情 >

针对个别店名经常识别错误情况，可外挂店名难例RAG库，embedding来自微调的1B模型。
对于店名字体模糊的问题，基于poi距离限制和店名相似度的模糊搜索通常可以匹配到目标店铺。

实际用例



query: 这边有没有乡村基啊?
直接基于query内容锁定到目标店铺。

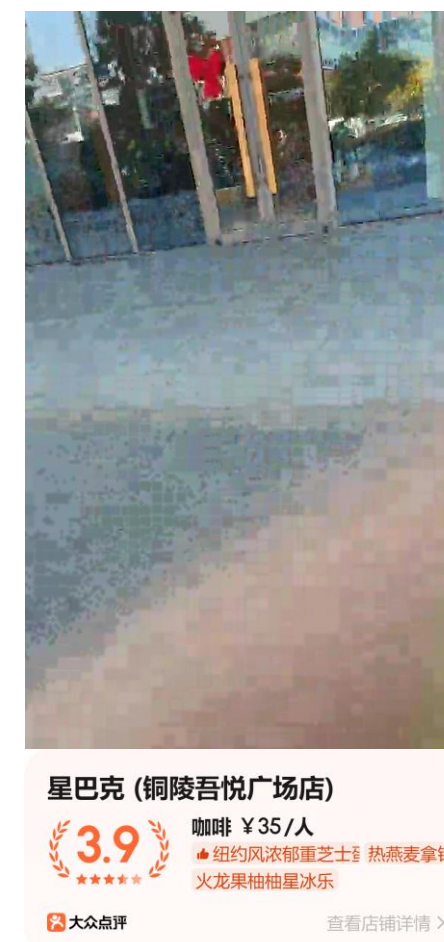


Turn1-query: 请帮我看一下这家店评分怎么样?

Turn1-answer: 这家STARBUCKS COFFEE的评分为3.9，整体评分一般，环境和服务也差不多。如果你有具体看法，记得告诉我，我可以帮你详细说一下。

Turn2-query: 牛逼

Turn2-answer: 哈哈，“牛逼”这个词用得很有劲儿嘛！不过星巴克（铜陵吾悦广场店）的评分是3.9，整体评价一般，环境和服务也差不多。如果你有具体想了解的地方，可以告诉我，我帮你详细说一下哦。



04 Citywalk未来演进

■ 产品技术

产品



- 1、结合室内定位的商铺识别，克服传统GPS定位不准导致的分店匹配错误。
- 2、从AI探店到AI评价，用户进入店内通过AI撰写评论反馈给三方平台，形成信息闭环。
- 3、实景主动推荐，在用户提问之前或走近商铺之前，结合用户画像主动为用户推荐餐厅。

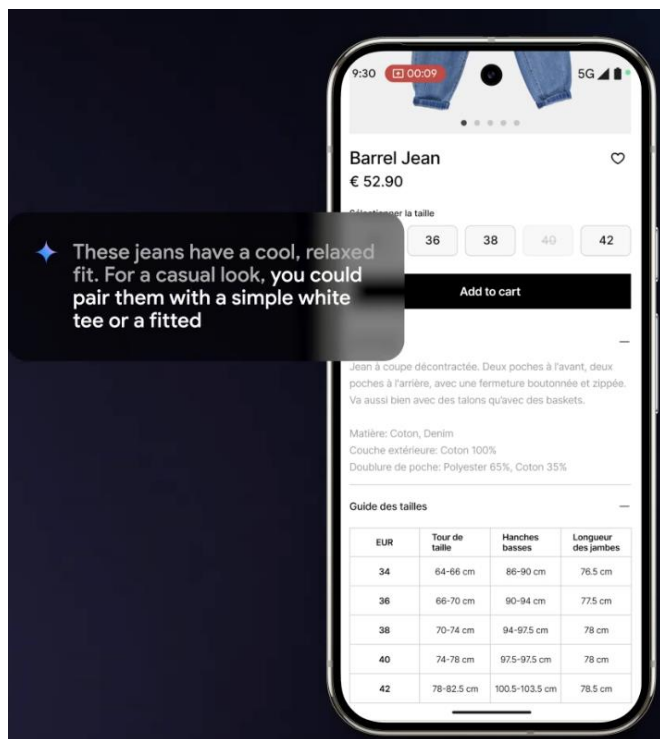
技术



- 1、Omni全模态模型应用，减少语音和视觉模态之间的错误传递。
- 2、跨平台Agent调度，例如餐饮agent和室内地图agent做联合信息总结。
- 3、自适应帧率和分辨率，根据场景难易自适应调整帧分辨率、结合画面相似度动态调节帧率，降低整体链路时延。

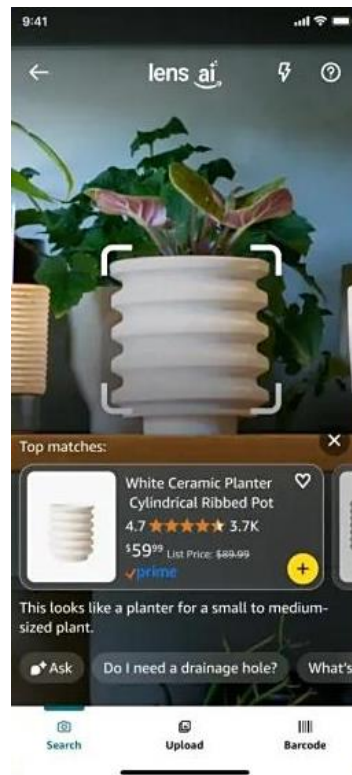
05 VideoAgent行业观察

Live AI



Google Gemini Live

“屏幕共享”功能，让用户与Gemini分享手机屏幕上的内容并询问有关问题。例如，一位用户正在购买一条宽松牛仔裤，并询问Gemini哪些其他衣服可以与之搭配。



Amazon Lens Live

许多消费者在实体零售店购物时，会对比同款或相似商品在亚马逊上的价格，寻找更优惠的选择。在使用Lens Live时，用户只需点击相机取景框中的任意物品，即可触发功能并聚焦该产品。

Others

1. Meta AI支持实时视觉问答，可通过Instagram相机与AI对话。
2. Canva AI Video Assistant视频生成+编辑助手，支持摄像头实时指令和屏幕协作。
3. Notion AI支持屏幕共享+摄像头互动，用于文档协作与AI问答。
4. Zoom AI Companion会议中可实时解读屏幕/摄像头内容，提供摘要与问答。
5. Snapchat My AI支持实时摄像头视觉互动，回答视觉相关内容。
6. Adobe Firefly Video AI视频生成+编辑助手，支持摄像头实时指令与屏幕协作。

极客邦科技 2026 年会议规划

促进软件开发及相关领域知识与创新的传播



参会咨询



查看会议



THANKS

探索 AI 应用边界

Explore the limits of AI applications

AiCon

全球人工智能开发与应用大会