

QCon

全球软件开发大会

从稳定到共情：对话式 AI 与硬件交互的技术探索

陈策

网易云信 资深音视频和智能硬件开发工程师



目录

01 智能硬件市场概况

02 网易云信对话式 AI 技术架构

03 硅基生命也有温度

04 典型案例

05 未来展望

📍 北京

QCon

全球软件开发大会

会议时间：4月10-12日

- 大模型赋能 AIOps
- 越挫越勇的大前端
- 云上业务架构演进
- 多模态大模型及应用
- 海外 AI 应用创新实践

📍 北京

AiCon

全球人工智能开发与应用大会

会议时间：6月27-28日

- 端侧智能
- AI Agent
- 多模态大模型
- 金融+大模型

📍 上海

QCon

全球软件开发大会

会议时间：10月23-25日

- AI Agent
- 大模型训练推理
- 端侧 AI
- 搜推深度融合
- 数智金融

4月

5月

6月

8月

10月

12月

📍 上海

AiCon

全球人工智能开发与应用大会

会议时间：5月23-24日

- 通用大模型
- AI Agent
- 垂直领域应用
- 模型可解释性

📍 深圳

AiCon

全球人工智能开发与应用大会

会议时间：8月22-23日

- 模型效率与部署
- 多模态大模型
- 大模型安全
- 智能硬件

📍 北京

AiCon

全球人工智能开发与应用大会

会议时间：12月19-20日

- 通用大模型
- 智能硬件
- LMOPs
- 具身智能

极客邦科技 2025 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

智能硬件市场概况

● 市场规模与需求：

全球出货量及家居、穿戴等细分领域占比，用户需求

2023

全球智能硬件（含智能家居、可穿戴、教育/养老/陪伴机器人等）出货量约 18 亿台，同比 +12%；其中智能家居占比 62%，可穿戴占比 28%，其它（玩具、机器人、车载配件）

2025

预计 2025 年总出货量突破 23 亿台，复合增速 9-10%，主要增量来自带屏音箱、陪伴/养老机器人、AI 学习硬件。
市场的增量主要来自于AI对传统硬件的升级

技术现状与痛点



30 秒里，用户从‘放摇滚’跳到‘娃的数学作业怎么办’——话题光速切换，设备却得先猜‘我到底是音乐 DJ，还是辅导老师’，于是每次都要重新唤醒、选模型、建立上下文，对话被切成一段段孤岛。

交互碎片化



上周才告诉设备‘我花粉过敏’，今天它依旧推荐‘周末去植物园’——用户画像一断电就归零，设备永远‘第一次认识你’。

记忆缺失



无论你说‘我失恋了’还是‘我升职了’，设备都回同一句‘抱歉，我没听懂’配同一颗平静蓝灯——声音无温度，灯光无颜色，情绪被机械复读。

情感响应同质化

实时对话式AI需要解决的问题



02

网易云信对话式 AI 技术架构

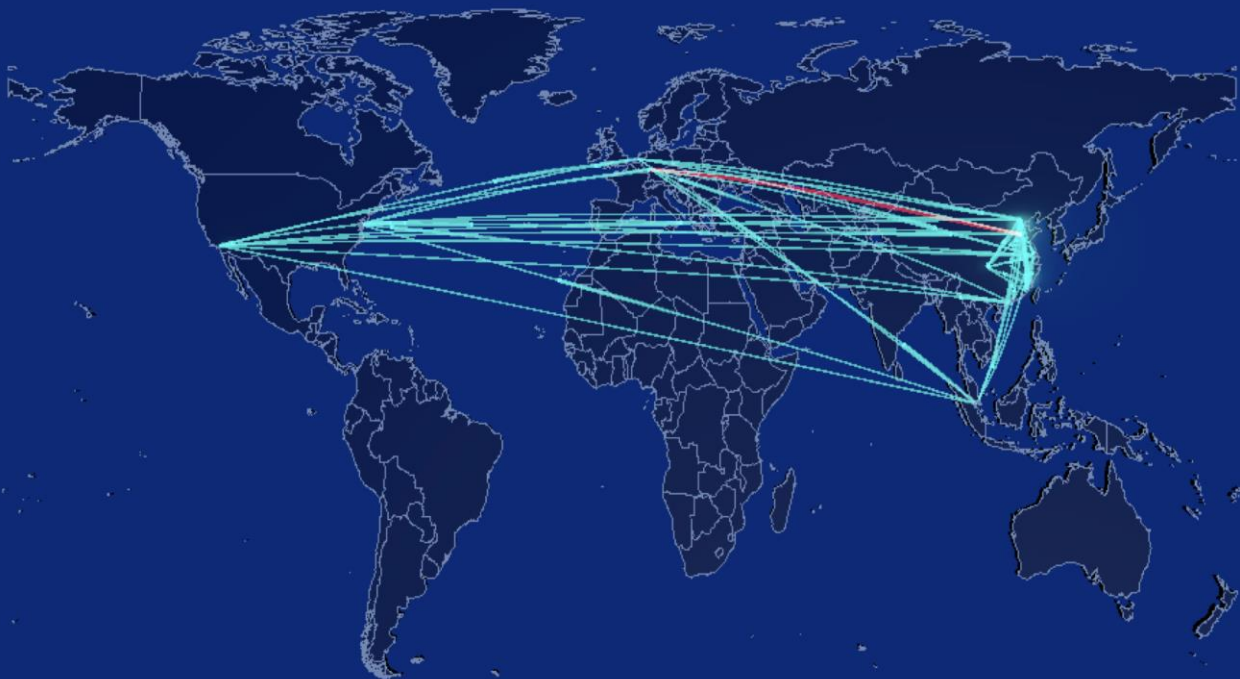
网易云信实时对话式AI智能体方案架构



- ✓ 客户端全终端RTC SDK
(iOS/AOS/Mac/Windows/Linux/小程序/Web/H5/RTOS/电话等)，实现跨端无缝对接
- ✓ 利用WE-CAN全球传输网络打通全球大模型，全流程流式处理，提供极致低延时
- ✓ 云端AI 音视频引擎，将云信多年的端侧音频处理能力应用在云端，提供优质的交互自然性和多模态体验
- ✓ 智能体平台支持个性化角色设定（Prompt、记忆、Tools、知识库...），提供个性化互动体验
- ✓ 全链路各AI 模块均支持动态插拔，适配用户自有AI 模块，降低接入门槛

网易云信RTC系统

全部			loss	avgRtt	jitter
起始城市	目的城市	出口运营商	目的运营商	搜索	
起始城市	目的城市	打分	lost	rtt	jitter
- 边缘服务器就近接入，缩短第一公里 - 全球加速网络WE-CAN - 音视频通话延迟最低200 MS					
		99	0.009	11	0.06
		99	0.007	31	0.08
		91	0.01	23	0.1
		99	0.003	26	0.06
		99	0.007	37	0.07
		99	0.009	38	0.05
		94	0.004	19	0.06
		99	0.003	36	0.09



- ✓ 客户端全终端RTC SDK（iOS/AOS/Mac/Windows/Linux/小程序/Web/H5/RTOS/电话等），实现跨端无缝对接
- ✓ 全球边缘接入，缩短第一公里。中间节点智能路由WE-CAN加速
- ✓ 音视频通话延迟最低至200 MS
- ✓ 丰富的音频算法积累，回声消除(AEC)，AI降噪等

AI-PipeLine

全链路流式 + 各模块极致的延迟优化



03

硅基生命也有温度

优雅打断

适配差异化场景的打断方案

实时双向对话场景

- 需要RTC来支持音频双向流式传输
- 需要具备AEC能力
 - 端侧AEC
 - 云端AEC

嘈杂环境

- 在一些环境嘈杂的场景下，自由语音打断容易出现频繁误打断。
- 对讲机式的交替对话。

明确的打断信号

明确“打断信号”，保留全双向流式对话“边听边说”的灵活性，避免无意义误打断



1

语音打断

自然对话打断
句首打断/句尾打断



2

手动打断



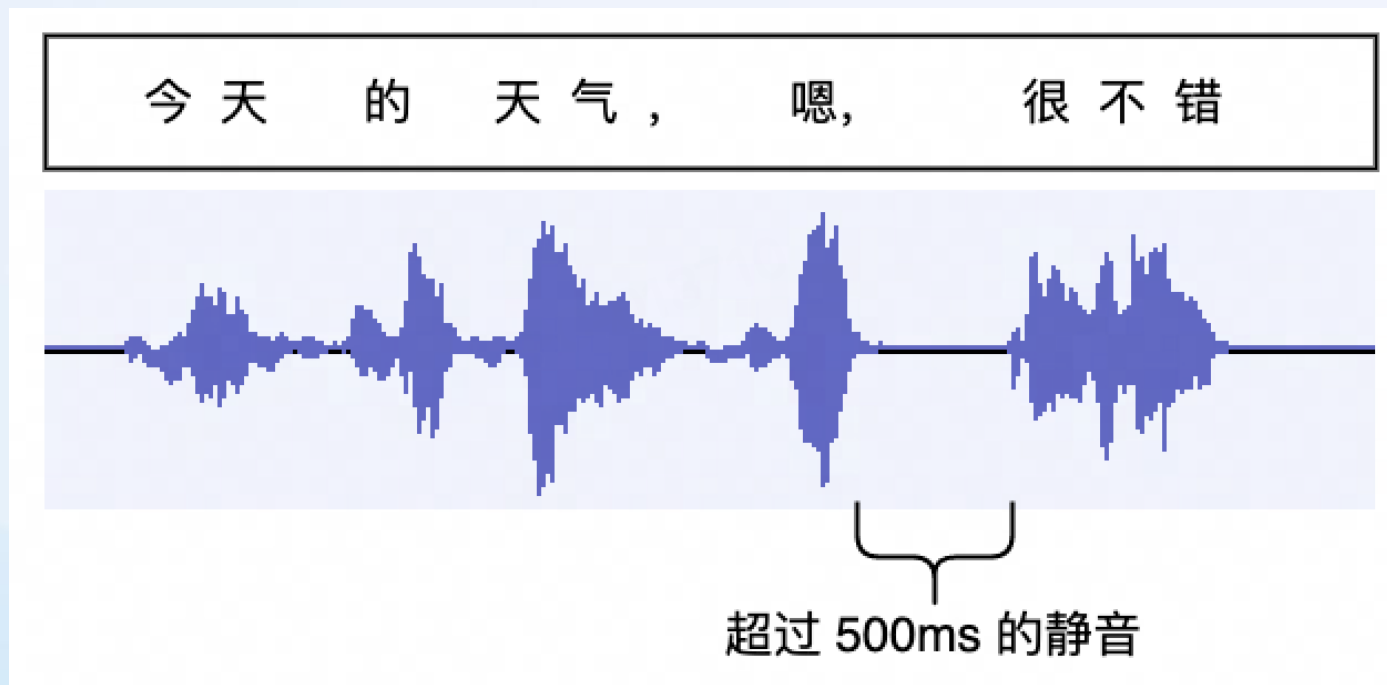
3

关键词打断

可自定义打断关键词

优雅打断

AI-Turn



用户说话被AI打断

VAD (Voice Activity Detection)

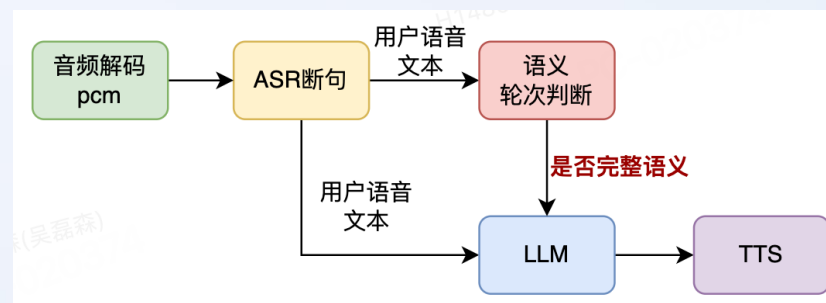
语音活性检测



被AI打断

(说“很不错”的时候，AI就开始回复了)

引入AI-Turn:



“今天的天气，嗯，很不错”判断是同一句话

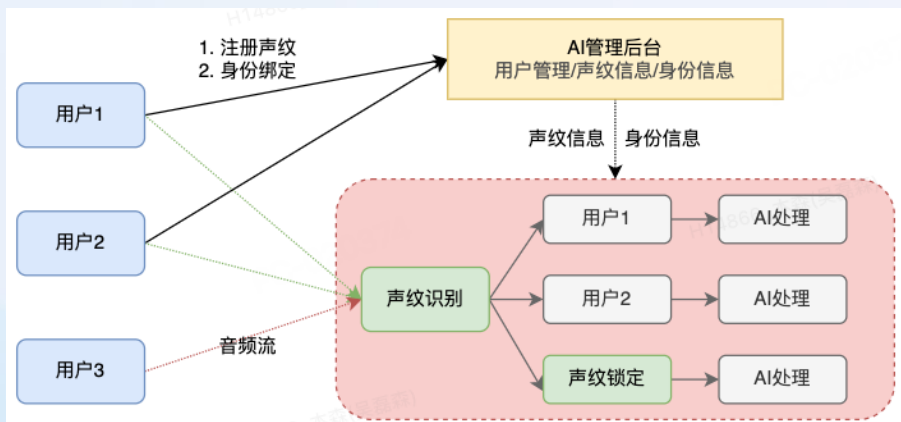
声纹识别

需求场景

- 区分身份，智能交互——家庭机器人，差异化智能体交互
- 专属唤醒，私密交互——AI陪伴，设备用户隐私保证

具体实现

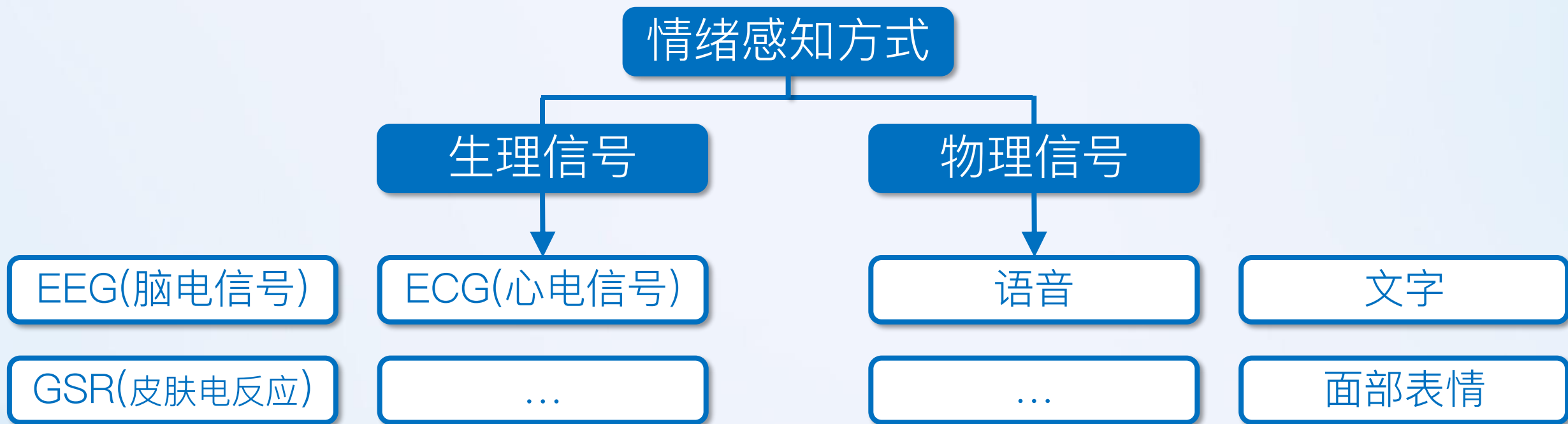
声纹注册和确认



- 声音录制 + 身份信息
- 独一无二的声纹标识
- 声纹识别，获取身份
- 可选择开启声纹锁定



情绪识别



- 语音信号及文字信号是便携式交互设备中最为广泛的信号形式，基于语音的情绪识别技术因而具有最广泛的应用范围。

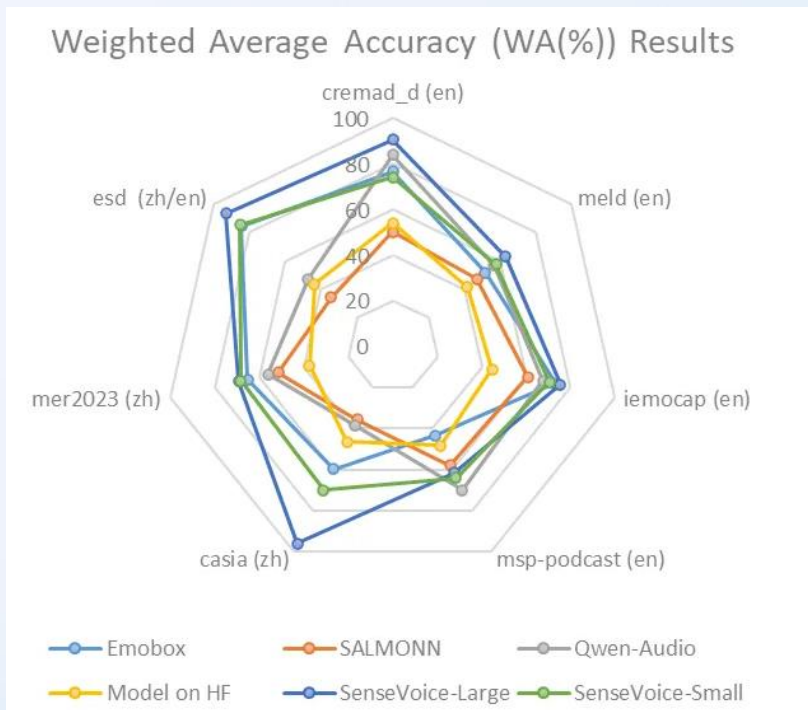
情绪识别

语音情绪识别(Speech Emotion Recognition, SER) 概况

- 国际: Whisper (Open AI) 、 HuBERT(Meta)、 Wav2Vec 2.0 (Meta)、 Data2Vec (Meta)
- 国内: SenseVoice (阿里) 、 Emobox (上交) 、 EMOTION2VEC (上交)
Vesper (华南理工) 、 WavLM (MSRA)
- GPT4o、 GPT4.5、 Llama 4等原生多模态大模型也具备情绪识别能力

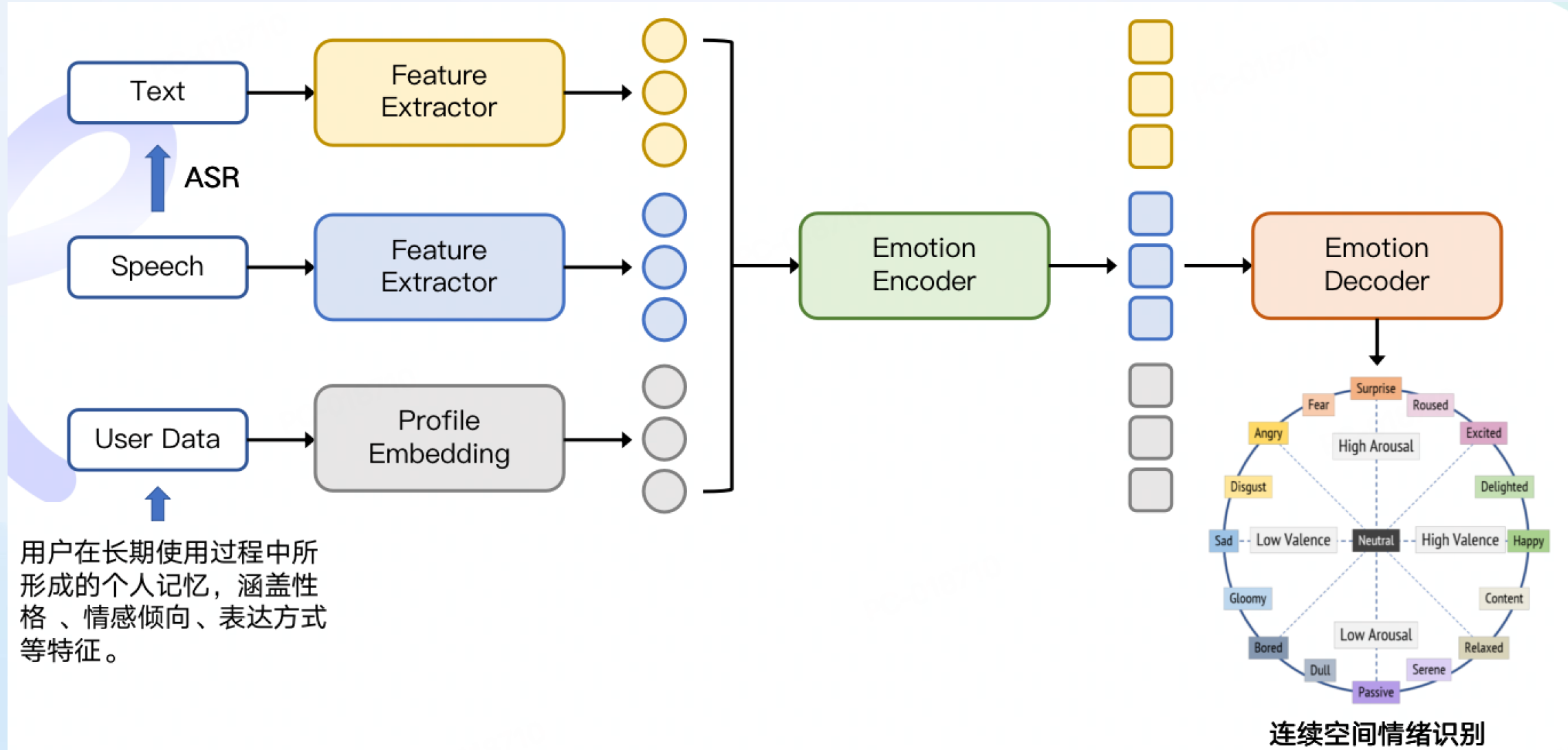
Model	Pre-training Corpus	Upstream	#Upstream Params	Downstream	#Downstream Params	WA(%) ↑
<i>Self-supervised Model</i>						
<i>small size</i>						
wav2vec (Schneider et al., 2019)	LS-960	Proposed	32.54M	Linear	0.13M	59.79
vq-wav2vec (Baevski et al., 2019)			34.15M		0.20M	58.24
<i>base size</i>						
wav2vec 2.0 (Baevski et al., 2020)	LS-960	Proposed	95.04M	Linear	0.20M	63.43
HuBERT (Hsu et al., 2021)	LS-960		94.68M		0.20M	64.92
WavLM (Chen et al., 2022)	LS-960		94.70M		0.20M	65.94
WavLM+ (Chen et al., 2022)	Mix-94k		94.70M		0.20M	67.98
data2vec (Baevski et al., 2022)	LS-960		93.75M		0.20M	67.38
data2vec 2.0 (Baevski et al., 2023)	LS-960		93.78M		0.20M	68.58
Vesper-4 (Chen et al., 2023a)	Mix-94k + LSED-206		63.52 M		0.26M	68.40
Vesper-12 (Chen et al., 2023a)	Mix-94k + LSED-206		164.29 M		0.26M	70.70
emotion2vec	LS-960 + Emo-262		93.79M		0.20M	71.79
emotion2vec*	LS-960 + Emo-262		93.79M		0.20M	74.48
emotion2vec	LS-960 + Emo-262		93.79M		0.20M	72.94
emotion2vec*	LS-960 + Emo-262		93.79M		0.20M	77.64
<i>large size</i>						
wav2vec 2.0 (Baevski et al., 2020)	LL-60k	Proposed	317.38M	Linear	0.26M	65.64
HuBERT (Hsu et al., 2021)	LL-60k		316.61M			67.62
WavLM (Chen et al., 2022)	Mix-94k		316.62M			70.03
<i>Supervised Model</i>						
TIM-Net (Ye et al., 2023)	-	MFCC	-	CNN(TIM-Net)	0.40M	68.29
MSTR (Li et al., 2023b)		HuBERT-large	316.61M	Transformer(MSTR)	27.00M	70.03
DST (Chen et al., 2023b)		WavLM-large	316.62M	Transformer(DST)	22.78M	71.80

不同模型在IEMOCAP情感数据集上对比结果



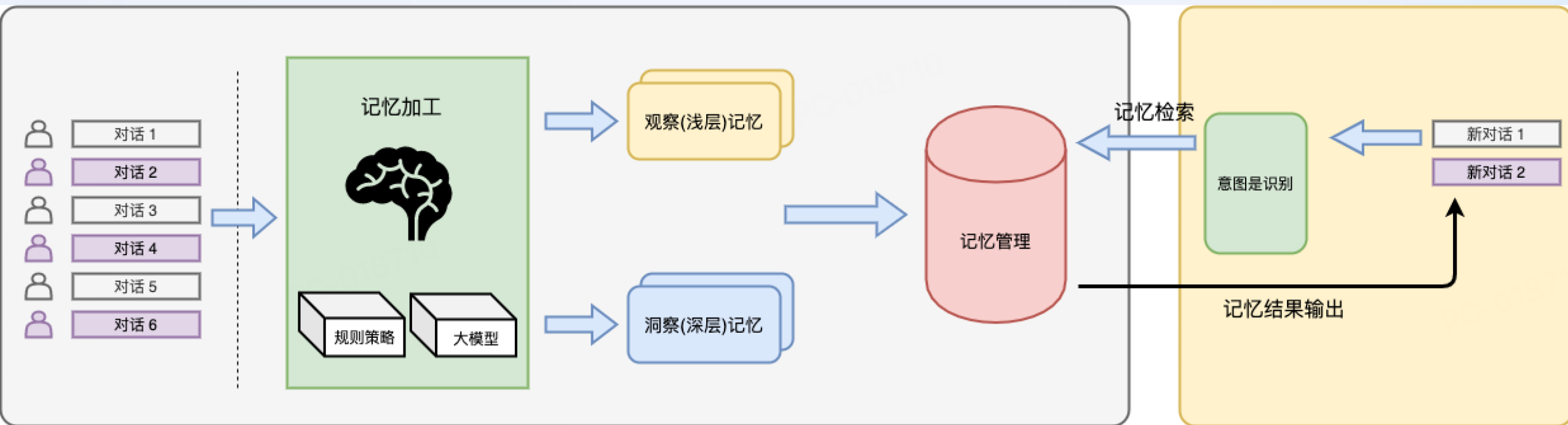
多模型、多数据集加权平均准确率对比

情绪识别



记忆与情感

记忆系统核心流程



记忆与情感

记忆分类



短期记忆 (STM)

即从对话流中获取的即时、未处理、基于会话界别的信息，短期记忆存储用户当前的交互行为和对话内容。人类的短期记忆的保留时间大概是几秒到十几秒，对于 AI 最有价值的短期记忆是近 10~20 轮对话的内容。

中期记忆 (MTM)

即从对话信息流中抽取的有价值的重要事实、事件和关系，中期记忆是短期记忆延续阶段、长期记忆的初始阶段，需要不断地、反复地抽取、总结、更新，保留记忆中最重要，最有价值的部分。

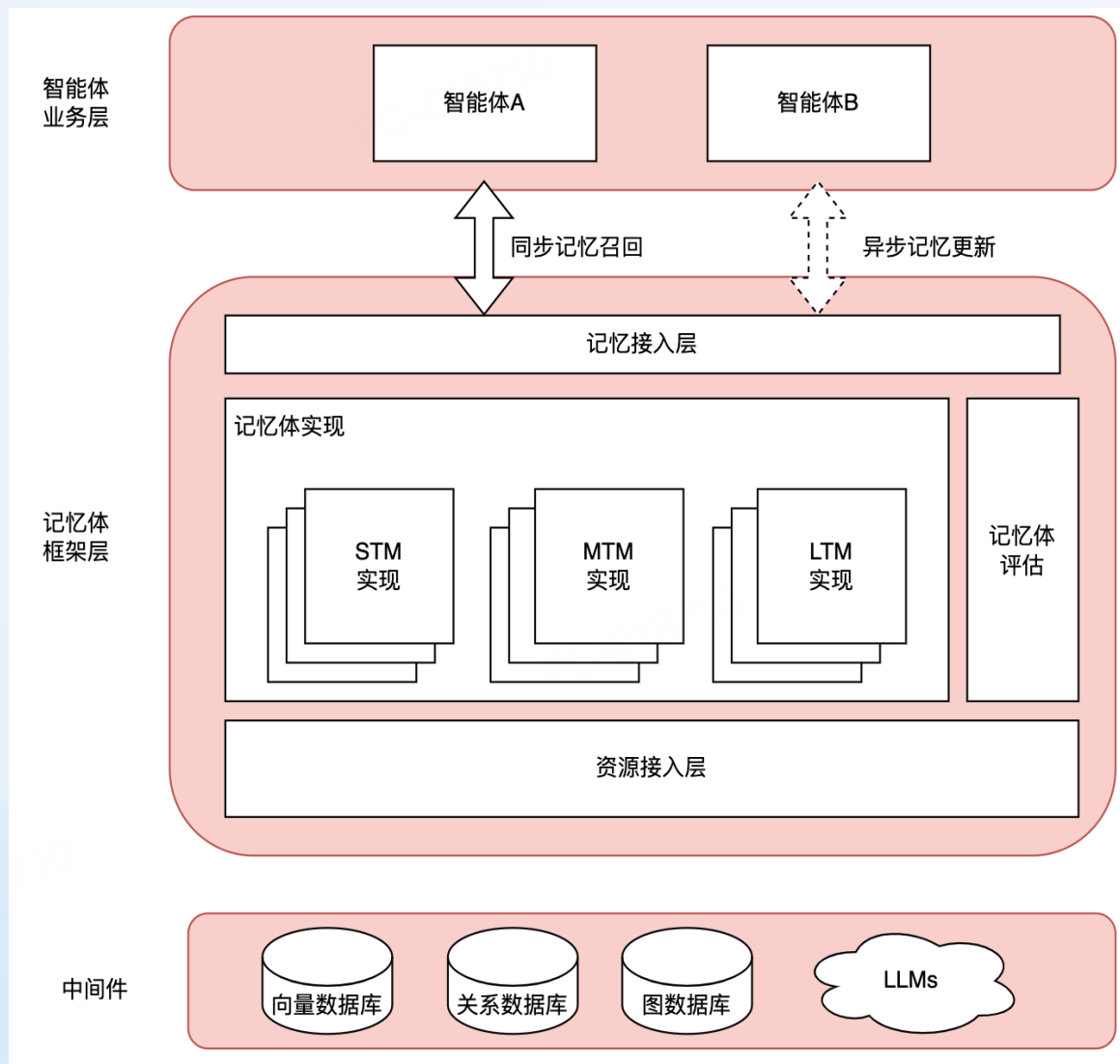
长期记忆 (LTM)

是通过短期记忆、中期记忆不断、深度加工而形成的持久的、稳定的信息。保留了个体知识、经验、技能和自我认同等核心信息。

预置记忆 (画像记忆)

可以给智能体预先注入一段记忆，比如医疗专业，用户画像，用户性格等。

记忆技术架构



• 用户的query:

今天下午，小华请我去吃炸鸡，你看要去吗？

• 记忆召回

短期记忆:

用户（昨晚）：工作了一天好累呀~~

AI（昨晚）：可以做一些简单的锻炼，放松放松。然后冲个热水澡

用户（昨晚）：嗯。坚持锻炼，保持身材

AI（昨晚）：加油

中期记忆:

用户制定了一个健身计划，需要控制饮食

用户喜欢吃炸鸡

用户和小华是好友

长期记忆:

用户很注重朋友关系，很nice。容易从众。喜欢聚会

• 期望AI的答复:

AI：要么拒绝掉呗。你最近在减少呢。昨天还说要保持身材哟。

AI：要么去吃呗。你们也很久没见面了。偶尔来一顿“欺骗餐”，也没关系。

AI：去吧~ 去吧~ 吃饱了，才有力气锻炼。

云信智能体配置平台

云信智能体配置平台



用户管理

基本信息、声纹信息、通讯录管理



Pipeline管理

打断模式、开场白、会话暂停、云端音频引擎配置



记忆模块

支持长期块状记忆、结构化记忆



角色配置

Prompt、LLM/ASR/TTS 配置



Tools/MCP管理

支持Tools工具链配置、支持添加MCP服务器



知识库配置

支持添加知识库

声音克隆

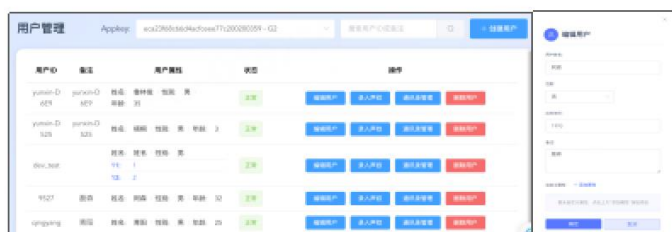
✓ 允许用户复刻独具个性化的交互音色。



✓ 用户

+

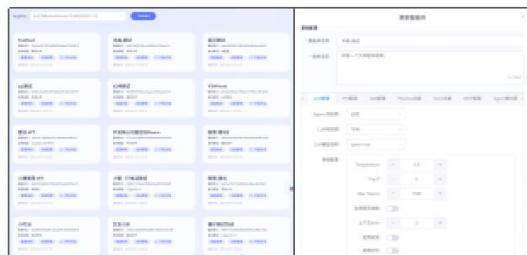
✓ 智能体
定义人设
匹配音色



【用户管理】



【用户声纹/通讯录信息】



【角色配置】



【Tools/MCP配置】



【知识库配置】



✓ 记忆

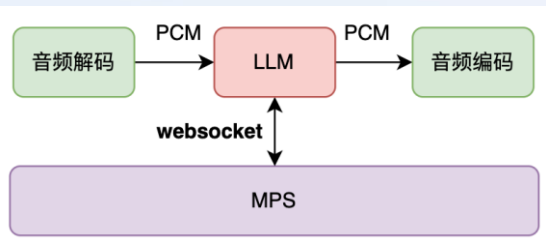
✓ 对话记录



多模态交互

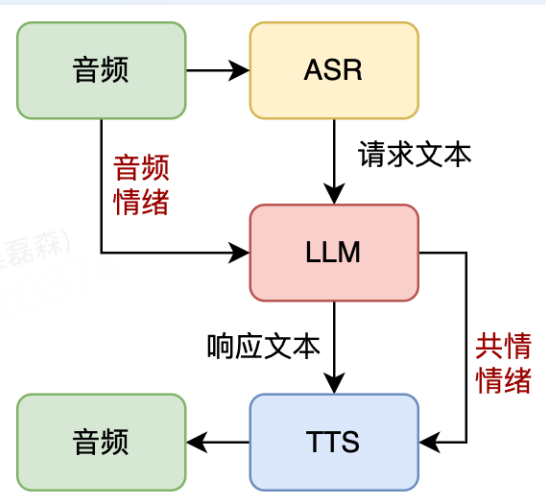
■ 音频多模态

□ RealTime



- ✓ 音频进，音频出
- ✓ 链路延迟降低至600ms以下
- ✓ 可控制输出音频

□ 情绪识别



- ✓ 20+基础情绪类型，允许自定义情绪库
- ✓ 情绪通过控制信令下发，控制终端



■ 视频多模态

□ 图片理解



支持终端APP（设备）通过控制信令通道直接上传图片，LLM完成对图片理解并做出响应

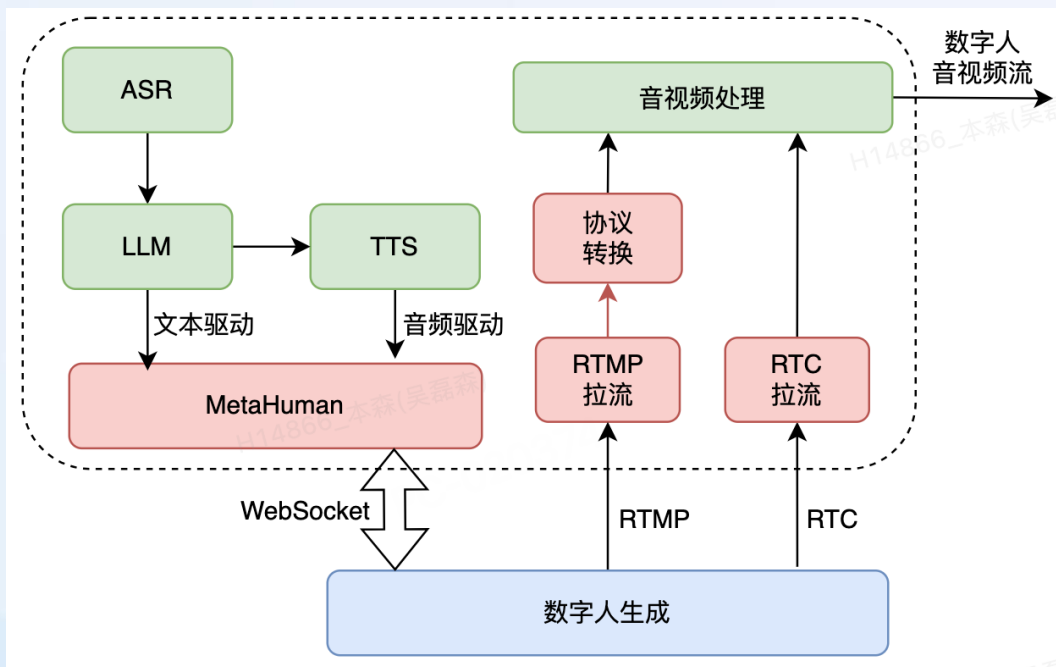
□ 视频流理解



AI-Pipeline订阅视频流结合CV能力，完成视频流理解
让设备具备视觉感知
<- 例如根据手势控制小车运动

多模态交互

数字人



- 支持文本驱动 / 音频驱动 2种方式
- 支持RTMP拉流 / RTC拉流



【聆听中】



【说话中】



【跳舞】

04 典型案例

陪伴类产品



【少儿陪伴：XX精灵】

主打儿童陪伴场景，给儿童讲故事，拍照识别，英语口语教学等。



【老人陪伴：XX智能】

主打老人陪伴场景，打电话，查天气，听新闻，跟APP互联互通



● 玩具产品



【AI闹钟】



【蛋仔派对玩偶】

情趣产品



【网易春风】

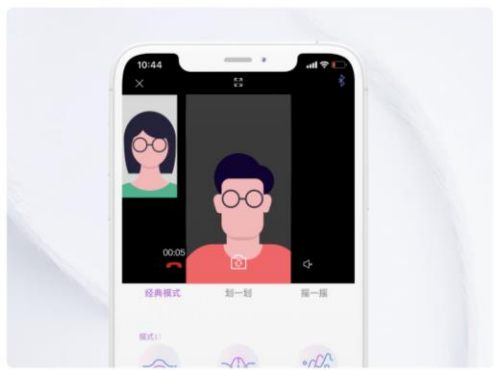
成人用品智能化方案

水平仪智能化方案



手机APP控制

蓝牙智能传输, 玩具一键掌控



互动模式

远程控制, 异地恋人也可实现音视频私密互娱



【XX科技】

云信AI能力展示



05 未来展望



未来展望

产品升级

原则上来说目前所有存量的物联网产品都有迭代升级的可能，只要降低硬件适配难度，智能硬件将肉眼可见的成为一大片蓝海市场。

生态构建

QCo致力于打造开放的多模态对话引擎生态系统，在芯片、硬件、app等领域吸引开发者和合作伙伴共同参与，构建内容安全、识别等技术闭环，促进技术的持续创新和应用的多样化发展，形成一个充满活力的对话式AI技术社区。

📍 北京

QCon

全球软件开发大会

会议时间：4月10-12日

- 大模型赋能 AIOps
- 越挫越勇的大前端
- 云上业务架构演进
- 多模态大模型及应用
- 海外 AI 应用创新实践

📍 北京

AiCon

全球人工智能开发与应用大会

会议时间：6月27-28日

- 端侧智能
- AI Agent
- 多模态大模型
- 金融+大模型

📍 上海

QCon

全球软件开发大会

会议时间：10月23-25日

- AI Agent
- 大模型训练推理
- 端侧 AI
- 搜推深度融合
- 数智金融

4月

5月

6月

8月

10月

12月

📍 上海

AiCon

全球人工智能开发与应用大会

会议时间：5月23-24日

- 通用大模型
- AI Agent
- 垂直领域应用
- 模型可解释性

📍 深圳

AiCon

全球人工智能开发与应用大会

会议时间：8月22-23日

- 模型效率与部署
- 多模态大模型
- 大模型安全
- 智能硬件

📍 北京

AiCon

全球人工智能开发与应用大会

会议时间：12月19-20日

- 通用大模型
- 智能硬件
- LMOPs
- 具身智能

极客邦科技 2025 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

大模型正在重新定义软件

Large Language Model Is Redefining The Software

