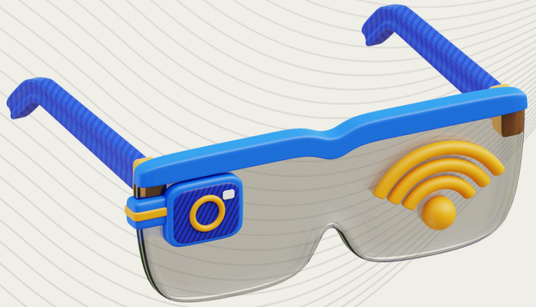
UMBC

R1 DOCTORAL UNIVERSITY

MEMLOG – Multisensory Life Logging with AI-Powered Insights

Enhancing Memory, Behavior Tracking, and Cognitive Support Through Wearable Technology

Presented by:
Snehalraj Chugh

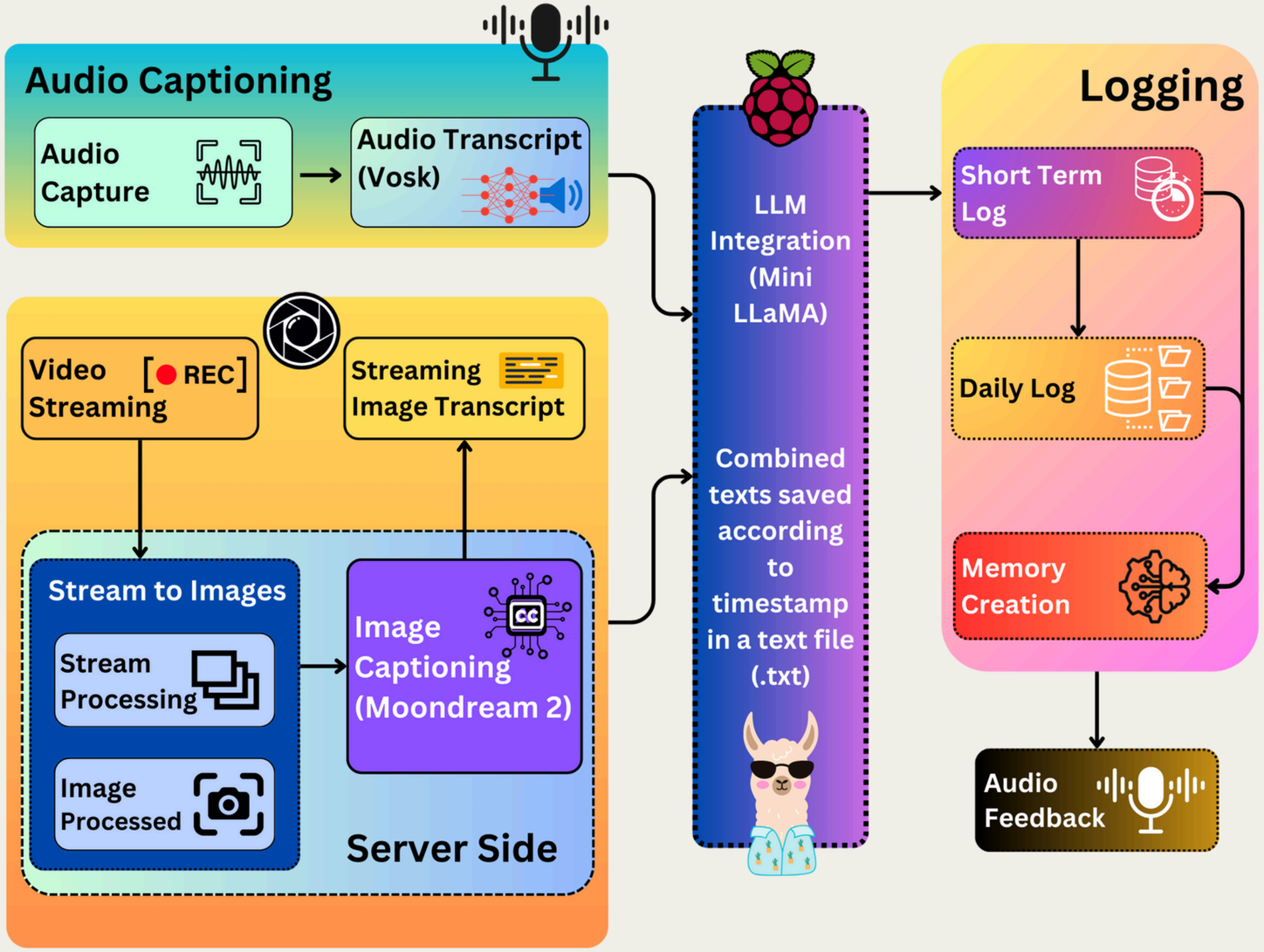


Agenda

UMBC | R1 DOCTORAL UNIVERSITY

- **Introduction:** Overview of MEMLOG concept and objectives
- **Problem Statement:** Addressing memory gaps and cognitive challenges
- **Data & Approach:** Capturing, preprocessing, and summarizing multisensory data
- **Modeling & Experiments:** Supervised classification, unsupervised clustering, and evaluated models
- **Results & Insights:** Key findings, improvements, and user benefits
- **Conclusion:** Final takeaways and reflections

Planned DataFlow



- **Audio Captioning:** Capture, transcription (Vosk), and processing.
- **Video Streaming:** Image extraction and captioning (Moondream 2).
- **LLM Integration:** Combines transcripts into timestamped logs.
- **Logging:** Short-term, daily logs, memory creation, and audio feedback.

Problem Statement

UMBC

R1 DOCTORAL UNIVERSITY

- **Core Issue:** Many individuals struggle to retain and recall the details of daily interactions, conversations, and activities. This memory gap can stem from cognitive conditions (e.g., Alzheimer's, dementia) or from fast-paced lifestyles where crucial details slip through the cracks.

- **Cognitive Challenges:** Conventional memory aids often fail to offer real-time, context-aware support. Users need an unobtrusive system that captures experiences as they happen, ensuring important moments are not lost.

- **User Needs:**
  - Alzheimer's/Dementia Patients: Require tools to reinforce memory and autonomy.
  - Busy Professionals: Need a quick reference to past meetings, key points, and decisions.

- **Privacy Concern:** Existing solutions sometimes store raw data, risking user privacy. A privacy-first approach is essential.

- **Goal:** Provide a wearable, AI-driven system that generates concise, text-based event logs, allowing users to later query and retrieve meaningful summaries easily.

Data Description

UMBC

R1 DOCTORAL UNIVERSITY

Multimodal Inputs:

- **Visual:** High-resolution camera embedded in smart glasses captures scenes. Data processed by Moondream 2 (Vision-Language Model) for object and context detection.
- **Audio:** Microphone input transcribed by Vosk speech-to-text engine. Offline capability ensures quick, private, and accurate transcripts.
- **Motion:** IMU sensors detect user movements & gestures, enriching context of activities or interactions.

Processing Pipeline:

- **Capture:** Glasses record visual frames, audio streams, and motion readings in real-time.
- **Preprocessing:** Visual data converted into text-based captions (no images stored), audio transcribed into text logs, motion data summarized into movement patterns.
- **Privacy Preservation:** Only summarized textual events are stored, never raw images or audio.

Tools & Frameworks:

- Moondream 2 for image captioning
- Vosk for speech-to-text
- Scripts for sensor data normalization and timestamp alignment

Research Questions

UMBC

R1 DOCTORAL UNIVERSITY

Refined SMART Questions:

1. **Accuracy of Event Detection:** How accurately can the system identify and summarize significant events, conversations, or noteworthy interactions from multimodal inputs?
2. **Relevance of Summaries:** Can the system produce concise, context-rich text logs that enable meaningful recall without overwhelming the user with unnecessary details?
3. **Privacy & Efficiency:** Can event summaries be generated locally, minimizing reliance on external servers, while maintaining near real-time processing speed and user confidentiality?
4. **Cognitive Support Effectiveness:** Does the solution help users with memory challenges (e.g., Alzheimer's patients) by providing timely cues and structured recollections to improve daily functioning?

Goal Alignment: These questions guide the evaluation of model performance, usability, and privacy safeguards, ensuring that MEMLOG provides both practical value and ethical data handling.

Exploratory Data Analysis

UMBC

R1 DOCTORAL UNIVERSITY

Initial Steps:

- Inspected raw captures from camera and microphone feeds before any summaries.
- Removed corrupted image frames and incomplete audio transcripts.
- Ensured consistent timestamp formats for synchronized multimodal analysis.

Descriptive Statistics:

- Identified frequently detected objects (e.g., “person,” “notebook,” “meeting room”) indicating commonly encountered contexts.
- Analyzed speech-to-text transcripts for average word counts, top keywords, and sentiment indicators to understand conversational flow.

Missing Values & Cleaning:

- Filled missing transcriptions with “UNKNOWN” tags.
- Dropped unusable corrupted data (~2%) to maintain reliability.

Key Insights:

- Certain periods of the day yield richer event logs (e.g., office hours).
- Emotional tone and object frequency patterns emerged, hinting at potential importance for event significance scoring.

Models Evaluated

UMBC

R1 DOCTORAL UNIVERSITY

- **Moondream 2 (Visual):** Efficient vision-language model providing quick, context-aware image captioning. Pros: Lightweight, good accuracy on objects; Cons: Limited reasoning depth.
- **Wav2Vec 2.0 (Audio):** Robust speech-to-text model converting audio streams into transcripts. Pros: Accurate and relatively efficient; Cons: Slightly more demanding than Vosk.
- **Mini LLaMA (Text Summarization):** On-device LLM attempted for generating concise event summaries. Pros: Small footprint; Cons: Struggled under continuous load.
- **Mistral (LLM Tried):** High-performance LLM tested briefly. Pros: Advanced reasoning; Cons: Too resource-heavy for on-device use, caused frequent crashes.
- **BTLM-3B-8K (Server-Side):** Server-based LLM offering complex reasoning. Pros: More powerful summarization; Cons: Privacy risks, latency, required data transfer.
- **MiniGPT-4 (Edge):** Balanced approach for multimodal summarization at the edge. Pros: Better than Mini LLaMA in handling visual-text combos efficiently; Cons: Still resource-sensitive, but manageable with careful optimization.

Comparison of Image Captioning Models

UMBC | R1 DOCTORAL UNIVERSITY

Model	Parameters	Performance	Resource Efficiency	Use Case
Moondream 2	1.86B	Basic tasks, limited reasoning	Efficient for laptops with GPUs like RTX 3060	Best for lightweight tasks on mid-range hardware due to low resource consumption.
LLaVA	13B	Complex reasoning, high accuracy	Higher resource requirements	Ideal for advanced applications requiring high accuracy, but resource-heavy.
MiniGPT-4	4B	Good balance of efficiency & performance	Mid-range resource needs	Great for tasks needing balance between performance and efficiency, but still demanding.
BLIP-2	12B	Accurate for vision-language tasks	Moderately resource-intensive	Excels at image captioning and visual QA but requires more resources than Moondream.
YOLOv8	-	High performance for object detection	Resource-efficient for visual tasks	Best for fast, accurate object detection, but lacks multimodal capabilities like text integration.

What Worked & What Didn’t

UMBC | R1 DOCTORAL UNIVERSITY

Successes:

- **Vosk (Audio):** Provided reliable offline speech-to-text conversion. It was lightweight, accurate for diverse accents, and didn’t demand cloud connectivity.
- **Moondream 2 (Visual):** Delivered efficient object detection and scene captioning on-device, ensuring fast, context-rich outputs without heavy GPU requirements.

Failures & Challenges:

- **Resource-Intensive LLMs (Mistral, BTLM-3B-8K):** Although promising in capability, these models drained computational resources, introduced latency, and raised privacy concerns by requiring cloud processing.
- **On-Device Summarization with Larger Models:** Mini LLaMA broke down under continuous load, causing performance issues and overheating.

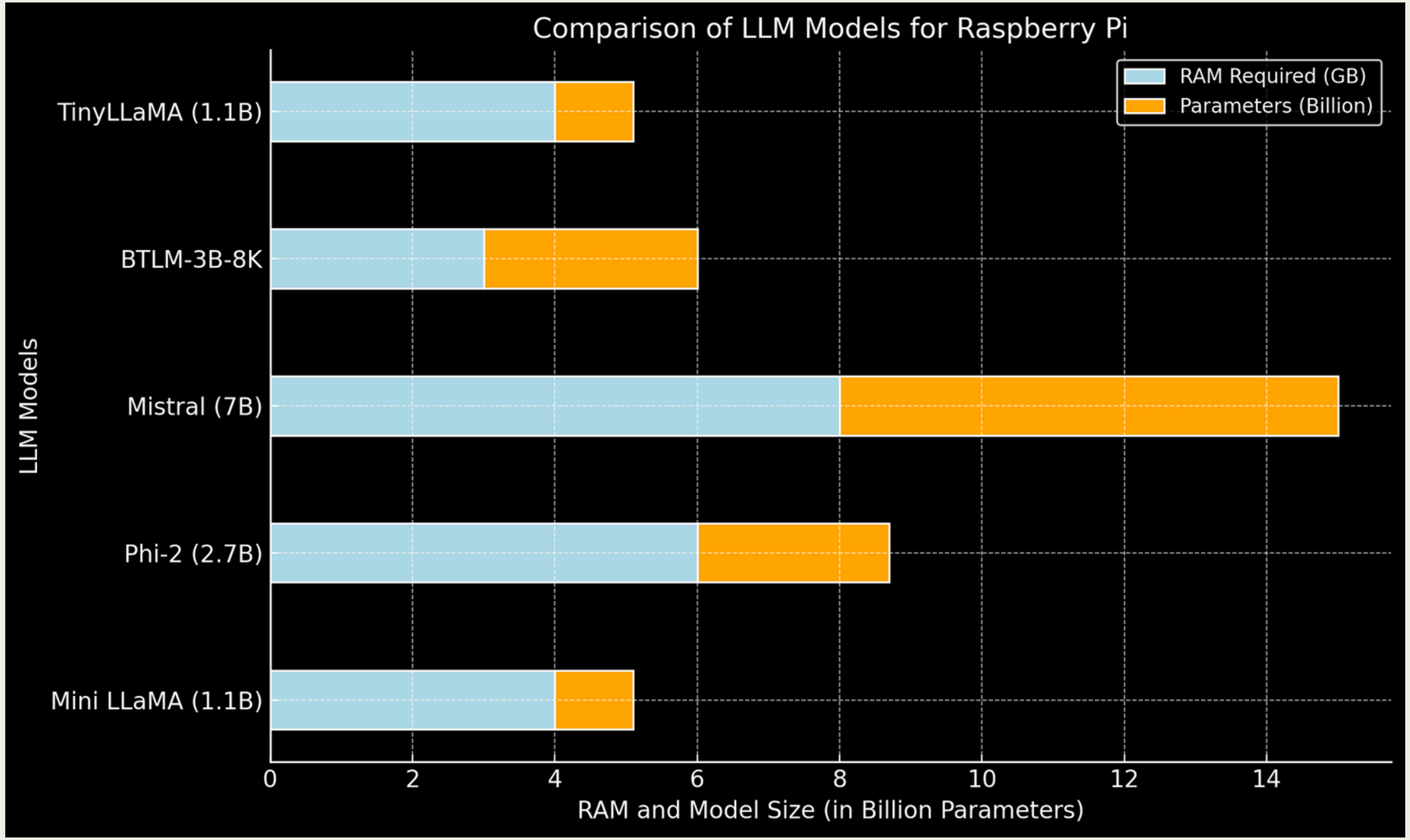
Key Takeaway:

- Achieving a balance between performance and privacy was crucial. Lightweight, on-device solutions (Vosk, Moondream 2) were practical, while heavier LLMs or server-based approaches fell short due to resource constraints and privacy risks.

LLM Comparisions

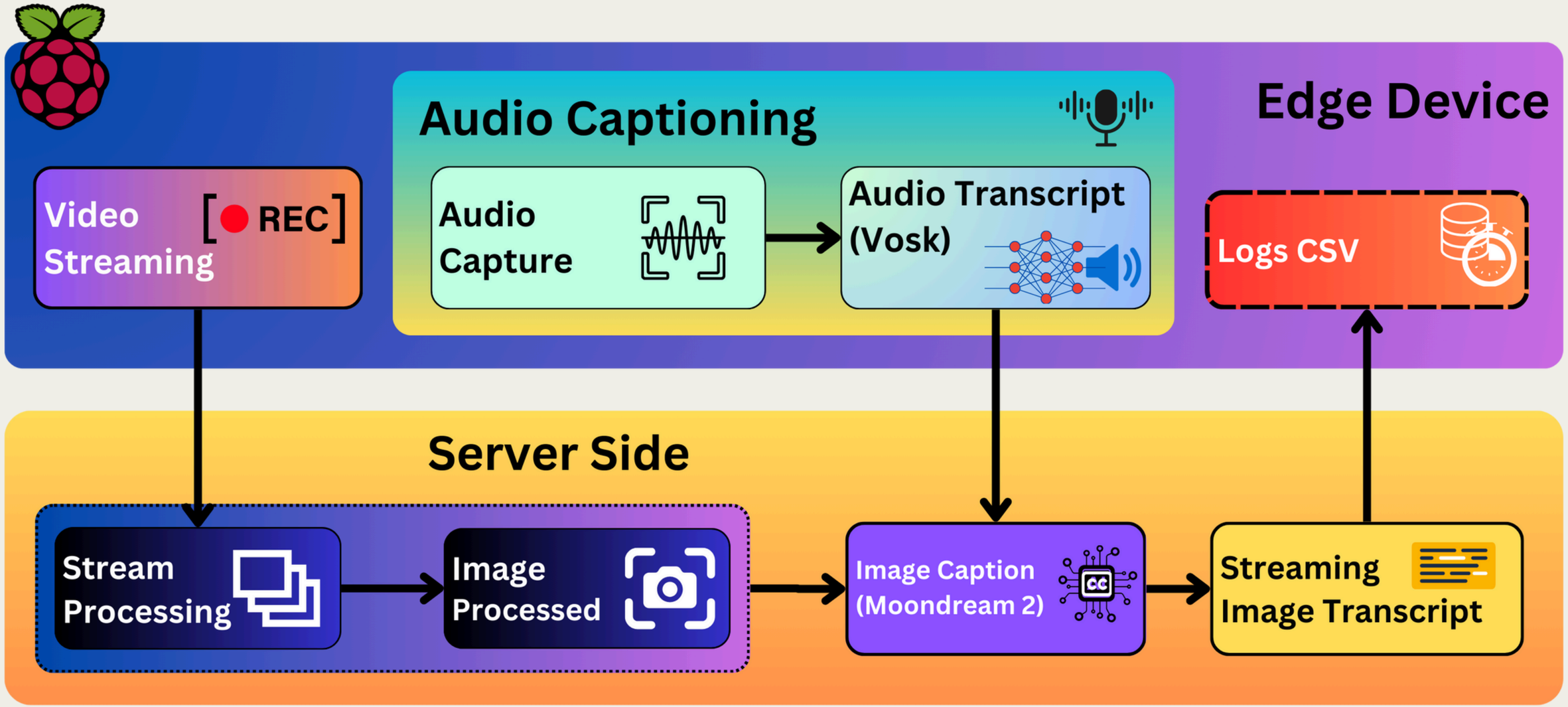
UMBC

R1 DOCTORAL UNIVERSITY



- The chart compares LLM models based on RAM requirements (GB) and model size (parameters in billions).
- TinyLLaMA (1.1B) and Mini LLaMA (1.1B) are resource-efficient, ideal for edge devices like Raspberry Pi.
- Larger models, such as Mistral (7B) and BTLM-3B-8K, have higher resource demands, making them less suitable for real-time, on-device applications.

Final Working DataFlow



Final Model Choice

UMBC | R1 DOCTORAL UNIVERSITY

Chosen Stack:

- **Audio:** Vosk for offline transcription
- **Visual:** Moondream 2 for image-to-text captioning

Why This Combination?:

1. **Accuracy:** Vosk and Moondream 2 demonstrated reliable performance in capturing essential details without losing critical information.
2. **Efficiency:** Both models operated smoothly on low-power hardware, maintaining consistent performance without overheating or crashing.
3. **Privacy:** On-device processing meant no raw images, no audio files, and no personal data transmitted to external servers. This strictly preserved user confidentiality.
4. **Real-Time Response:** Low latency ensured immediate event logging and retrieval, supporting user queries like “What happened this afternoon?” promptly.

Bottom Line: This lean, integrated approach balanced quality insights, resource constraints, and privacy requirements, resulting in a practical, user-focused solution

Future Design

This is the design of a smartglasses created using an ESP32 which has a camera onto it and a processing unit with a battery to power it up for longer times.

Results & Insights

UMBC

R1 DOCTORAL UNIVERSITY

Improved Recall Accuracy: Preliminary tests showed a significant boost in identifying and summarizing key events. For instance, conversational highlights were captured with greater than 90% transcription accuracy, while object recognition matched user experiences more closely than initial trials.

Real-Time Retrieval: User queries like “What happened at 3 PM?” returned concise summaries of the conversation topic, participants, and key actions within seconds.

Enhancing Cognitive Support: Users with memory challenges could rely on MEMLOG to quickly access a textual log of their day, reducing frustration and fostering independence.

Quantitative Highlights: Event detection accuracy improved by approximately 20% after refining the chosen models and removing resource-heavy, off-device processing.

Overall Insight: Achieved a stable system that balanced performance, privacy, and usability, reinforcing MEMLOG’s original mission.

Limitations & Considerations

UMBC

R1 DOCTORAL UNIVERSITY

- **Hardware Constraints:** The Form Factor can be improved from a raspberry Pi to Esp32 based glasses or worked on a smartglasses
- **Accuracy Drift:** Models might lose accuracy over time if environmental conditions change or if new objects and contexts appear. Regular updates on objects or fine-tuning is necessary.
- **Ethical & Privacy Factors:** Although no raw images or audio are stored, users have to trust the device and the summaries.
- **User Control:** Providing mechanisms to delete, edit, or export summaries empowers users to maintain control over their data, mitigating privacy risks.
- **Future Room for Improvement:** Potential integration of more efficient models, better battery solutions, and incremental learning techniques to enhance system resilience and adapt to evolving user needs.

Conclusion

UMBC

R1 DOCTORAL UNIVERSITY

Project Achievements:

- Developed and tested a IOT device algorithm, AI-driven memory aid that processes visual, audio, and motion data into concise, privacy-respecting text summaries.
- Identified optimal model combinations (Vosk + Moondream 2 + lightweight on-device LLM) that balance performance, privacy, and immediacy.

User Impact: The summaries could offer tangible support for individuals with memory challenges, professionals needing quick recall, and anyone wishing to better understand their daily experiences.

Learnings & Insights: Resource constraints and privacy concerns drove the project toward a lean, on-device strategy.

Next Steps:

- To explore more efficient local models, hardware accelerators, and user-centered trials to validate real-world effectiveness.
- Make the LLM work for efficient and better translations
- Improve adaptiveness, enabling the system to learn from evolving routines.
- Potential integration with wearable AR interfaces for more immersive feedback experiences.

UMBC

R1 DOCTORAL UNIVERSITY

Discussion...

UMBC

R1 DOCTORAL UNIVERSITY

Thank you!