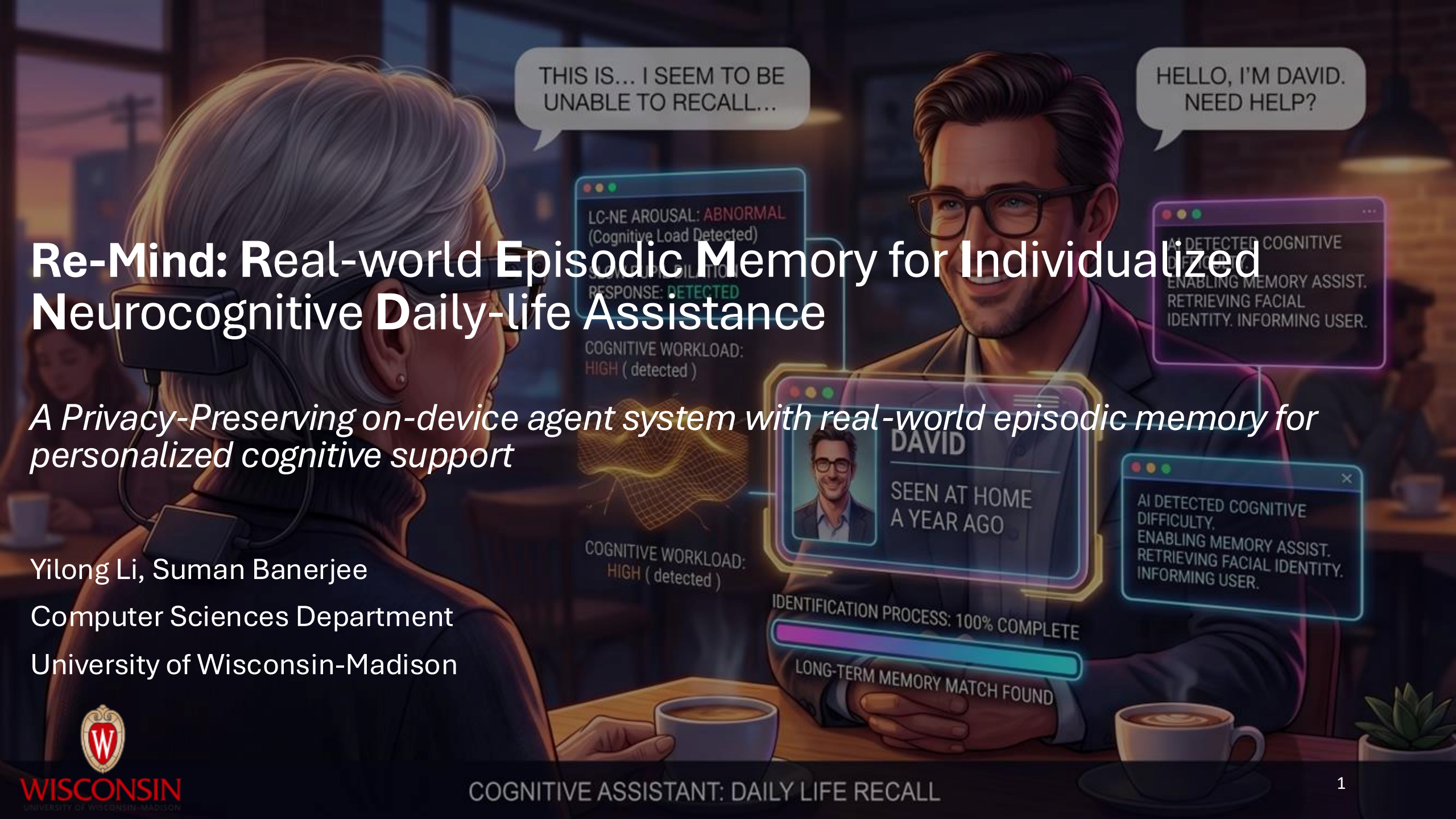
Re-Mind: **R**eal-world **E**pisodic **M**emory for **I**ndividualized **N**eurocognitive **D**aily-life Assistance

A Privacy-Preserving on-device agent system with real-world episodic memory for personalized cognitive support

Yilong Li, Suman Banerjee

Computer Sciences Department

University of Wisconsin-Madison

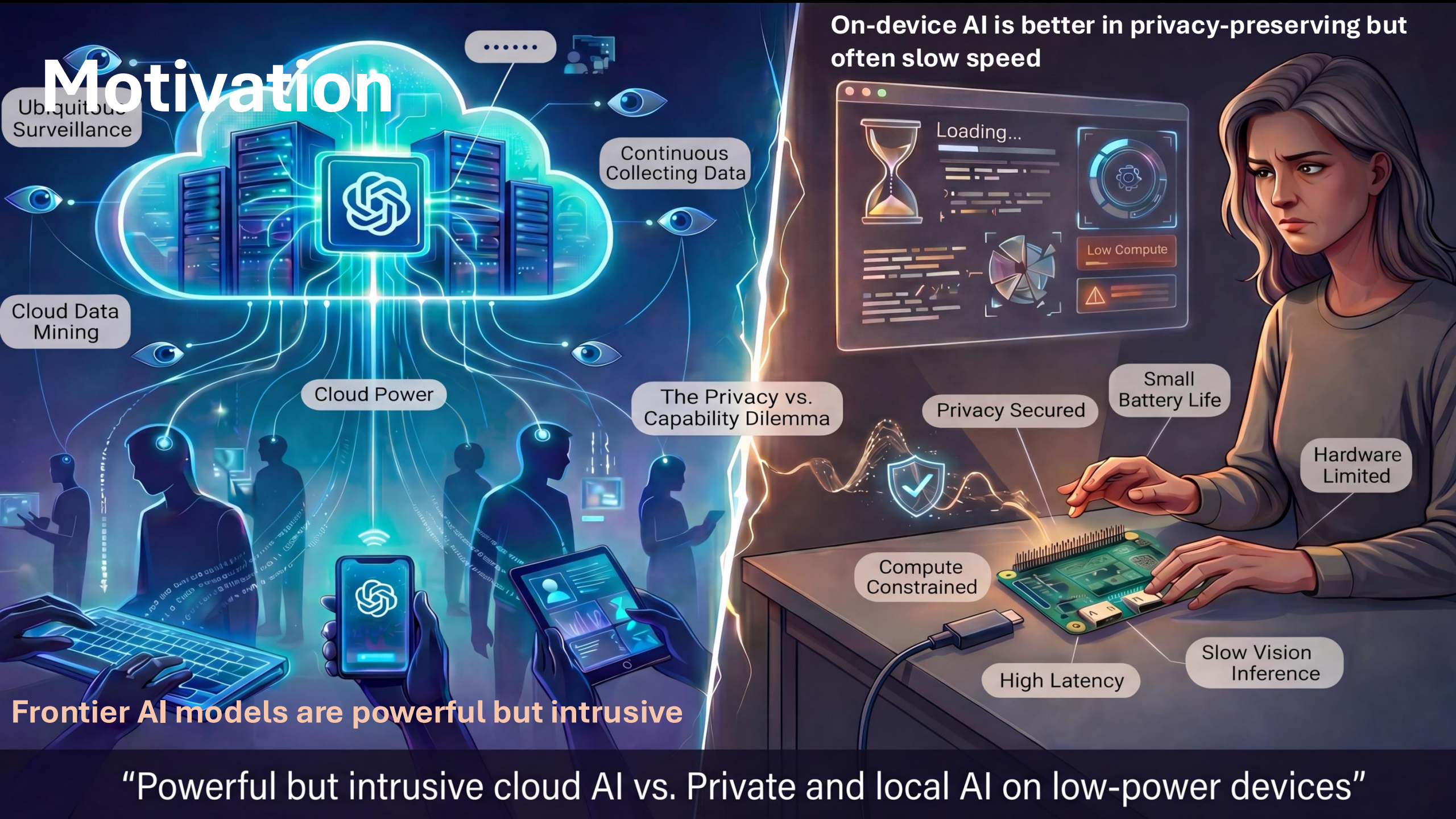


Motivation

AI Models and Agents are now powerful, but still lacking connection to people

"Powerful, but still behind the screen"

Motivation



"Powerful but intrusive cloud AI vs. Private and local AI on low-power devices"

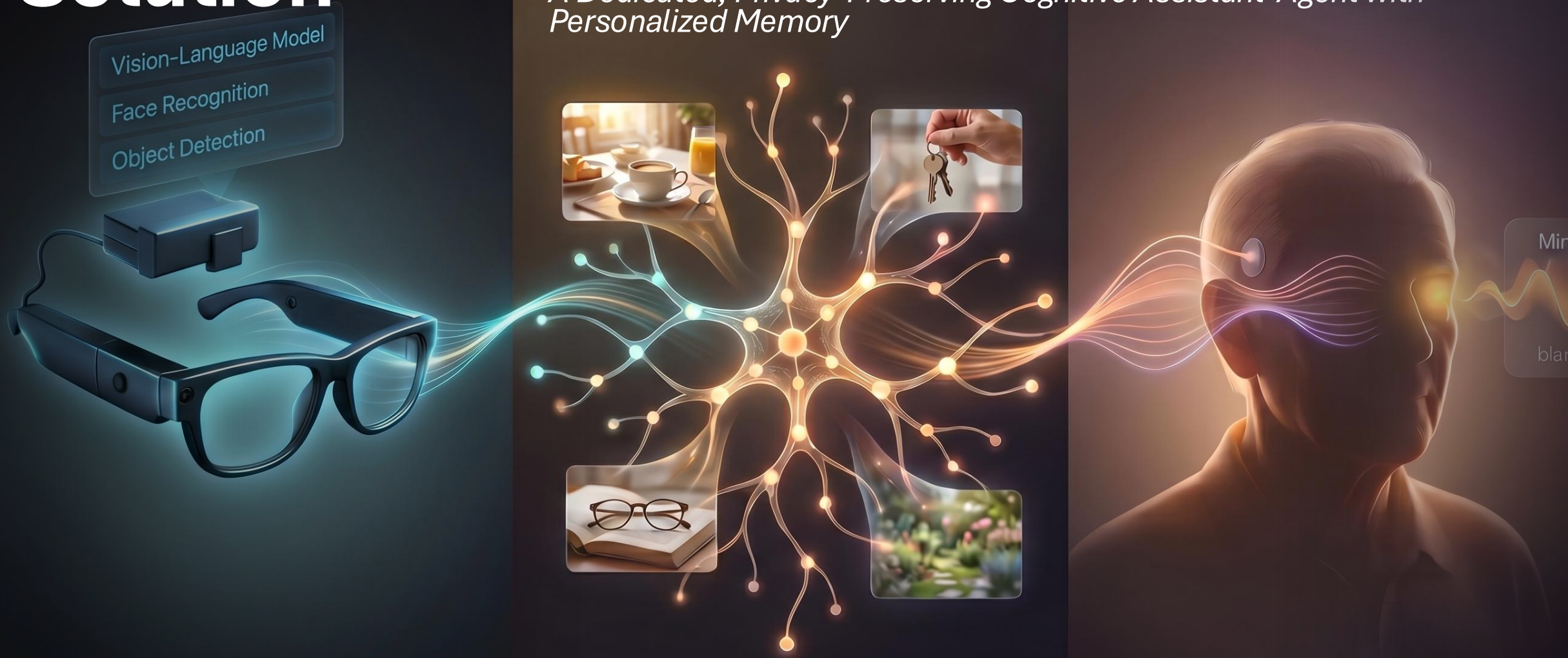
Motivation

Existing LLM Agents have no reliable memory

Solution

What We Build

A Dedicated, Privacy-Preserving Cognitive Assistant Agent with Personalized Memory



Efficient On-Device AI

Privacy-Preserving Personal Memory

Cognitive-State-Aware Assistance

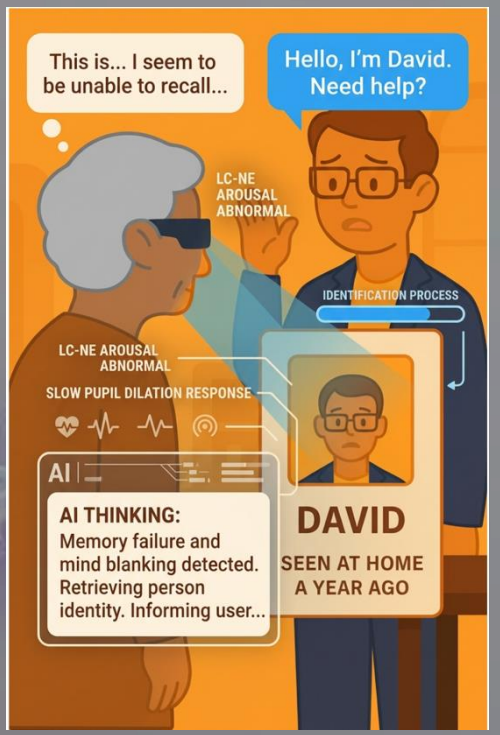
Solution

What We Want:

A Device for People's Cognitive Assistant and Accessibility

Visual Understanding

Find Item



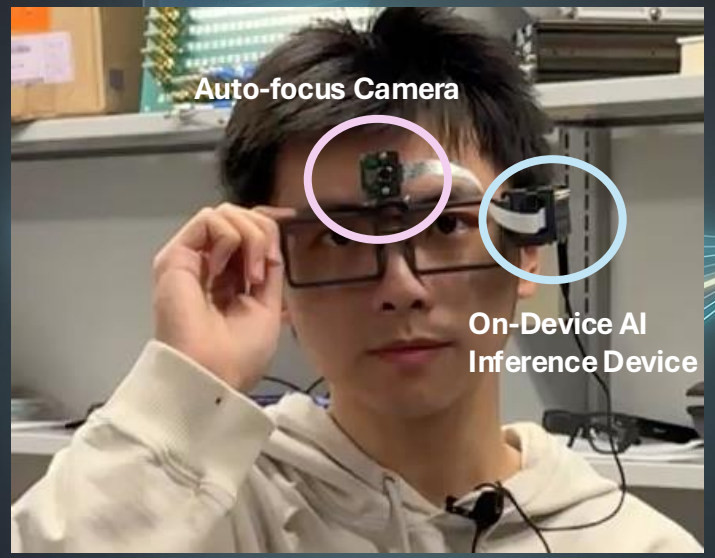
Memory-failure Awareness

Not Another Platform. The Right Device.

Solution

Solution

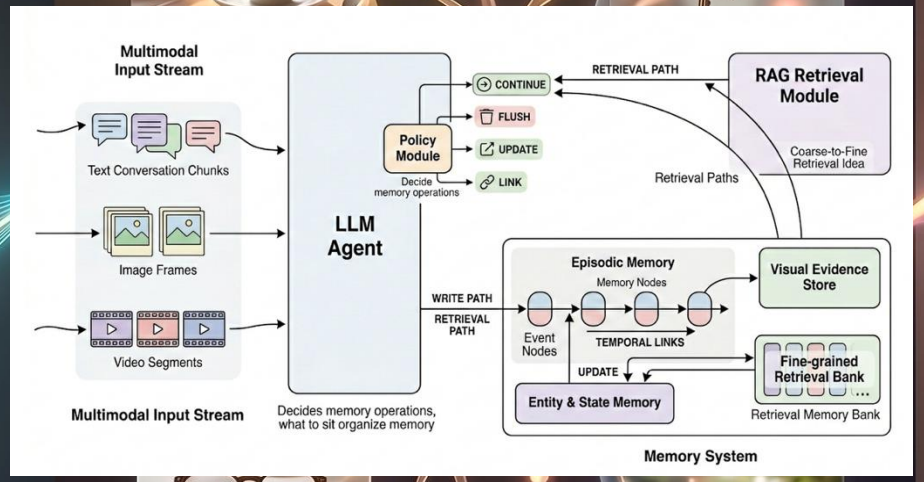
Vision-Language Model
Face Recognition
Object Detection

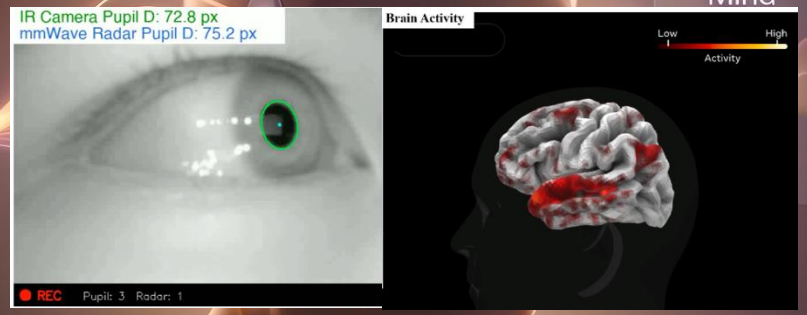


What We Have Now

A Real-Time Demo of Visual Understanding and Face-and-Item Memory on Our Custom-Built Hardware

Developed Techniques for Cognitive-State Estimation and Memory-Failure Awareness





Platform of Efficient Vision Language Model Inference on a battery-powered device

Efficient On-Device AI

Memoria: A privacy-preserving, continuously evolving personal agentic memory

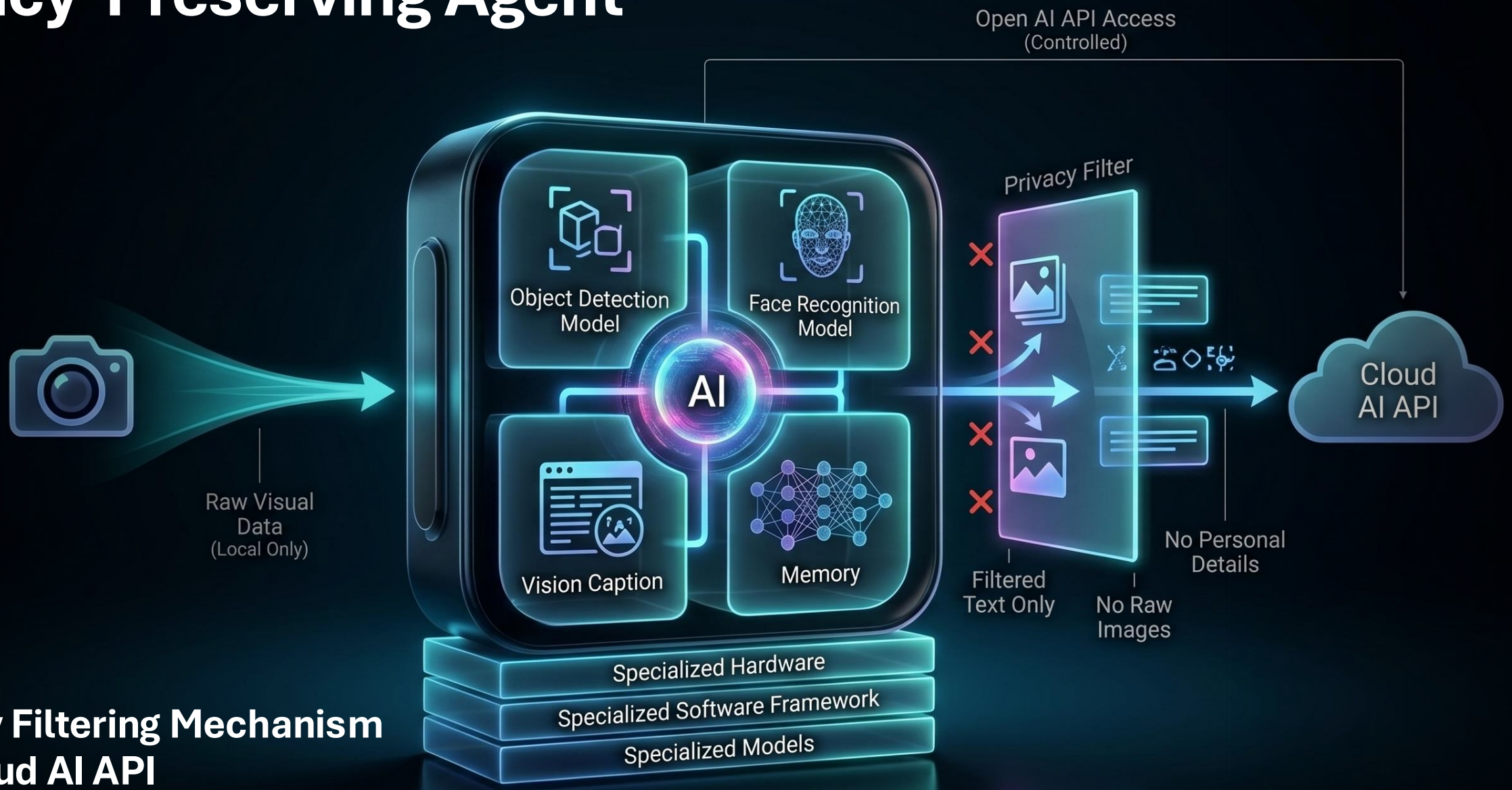
Privacy-Preserving Personal Memory

Detecting Memory Failure and Mind Blanking via LC-Linked Pupillary Responses

Cognitive-State-Aware Assistance

Privacy-Preserving Agent

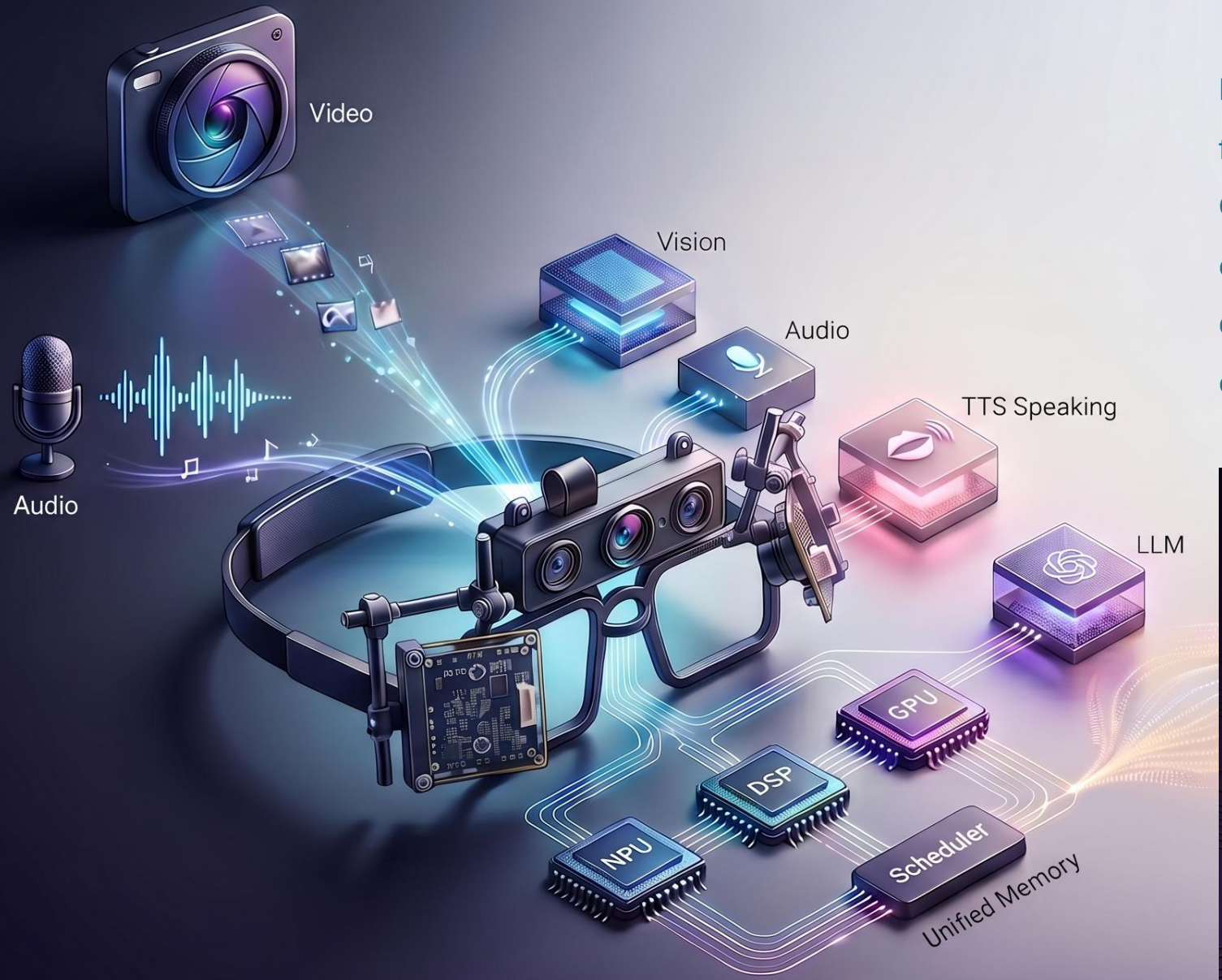
Privacy-Preserving Agent



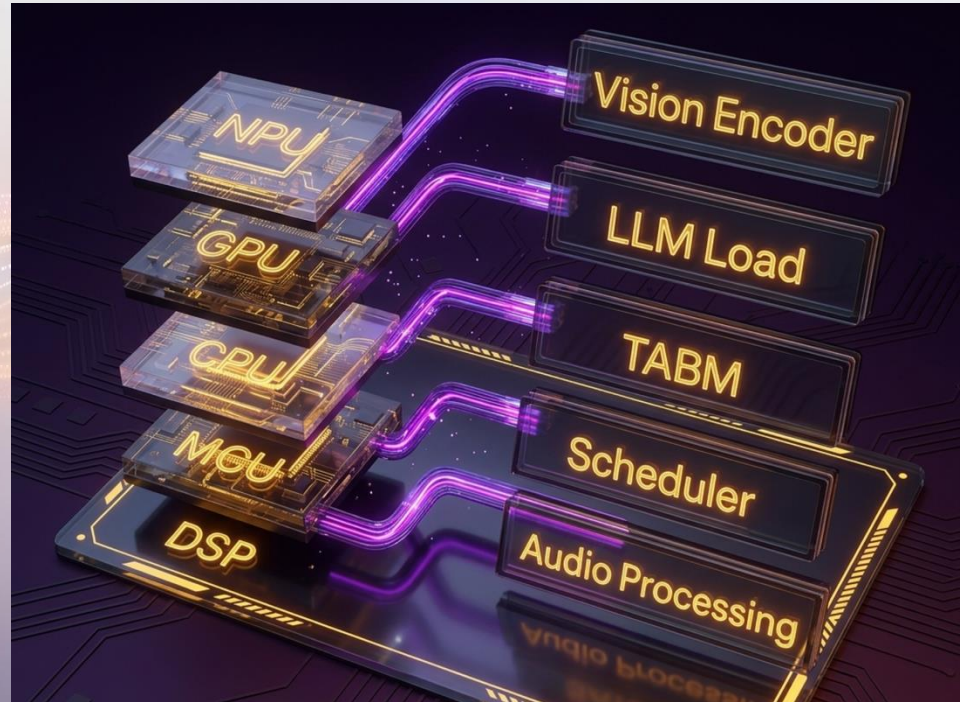
Privacy Filtering Mechanism for Cloud AI API

"Not Another Platform. The Right Device."

Efficient Cross-Accelerator Inference



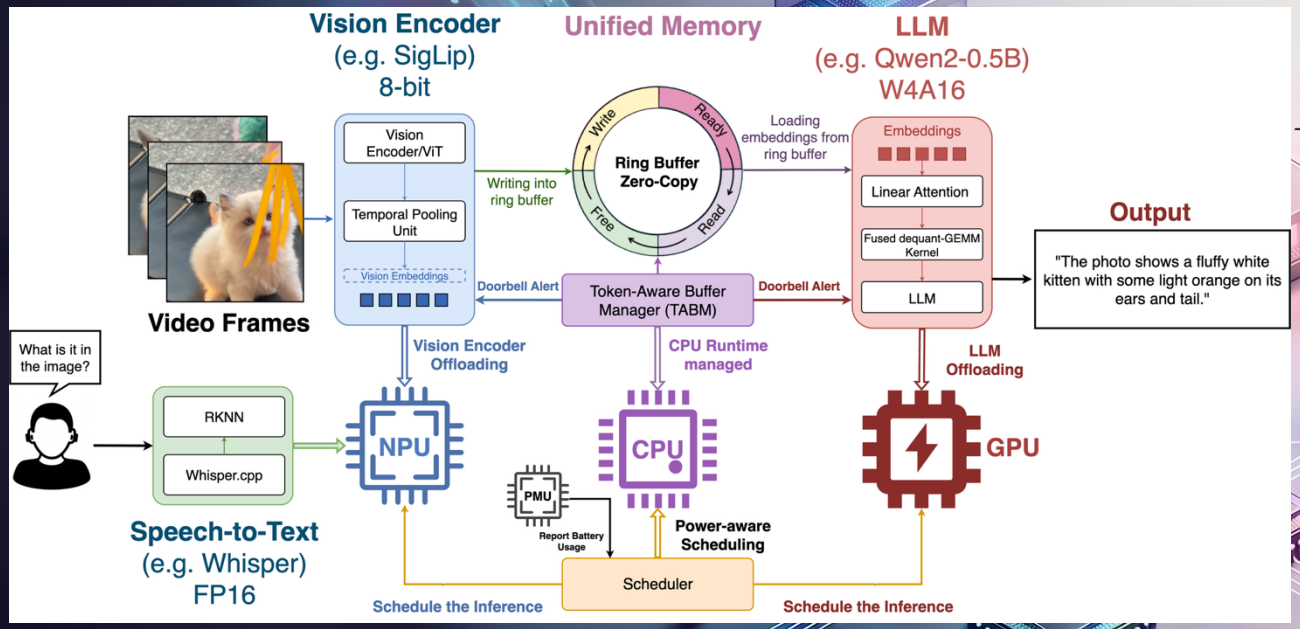
Many companies are building expensive custom AI chips for wearables. **In contrast, we rely only on low-cost COTS and readily available components.** By adopting a cross-accelerator design, we partition AI models across different cheap accelerators and connect them through our specialized memory architecture.



Efficient Cross-Accelerator Inference

Video

Vision

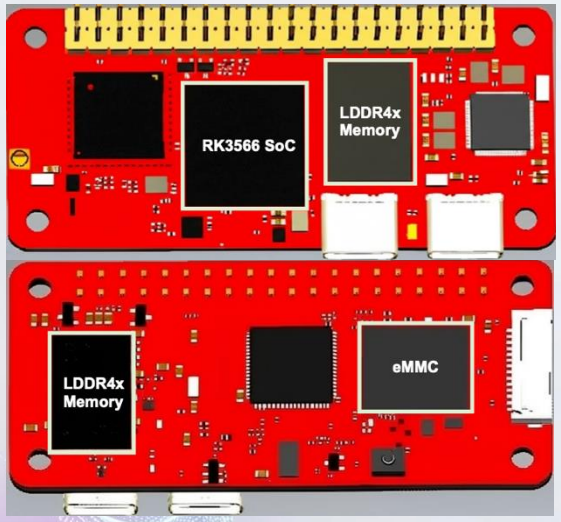


Use more accelerators on the SoC

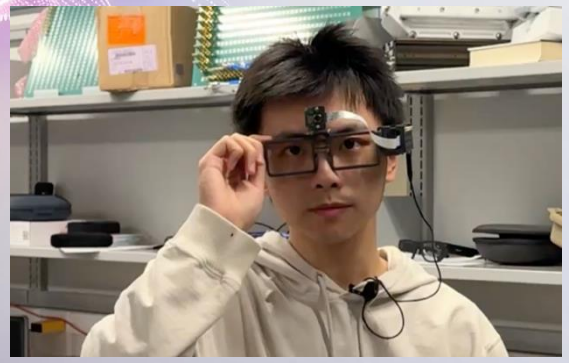
TTS Speaking

LLM

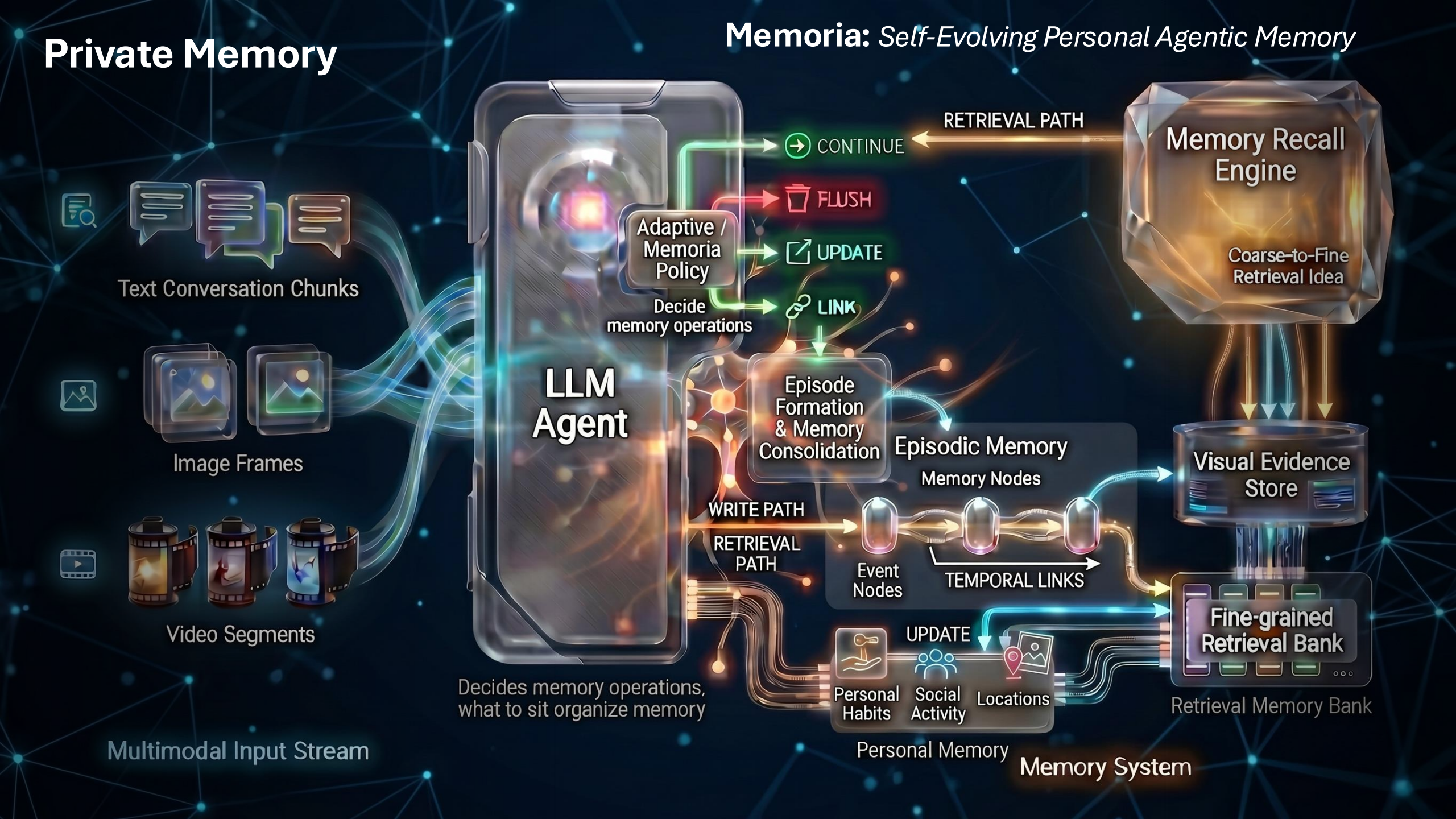
Unified Memory



Custom Low-Power Hardware Design

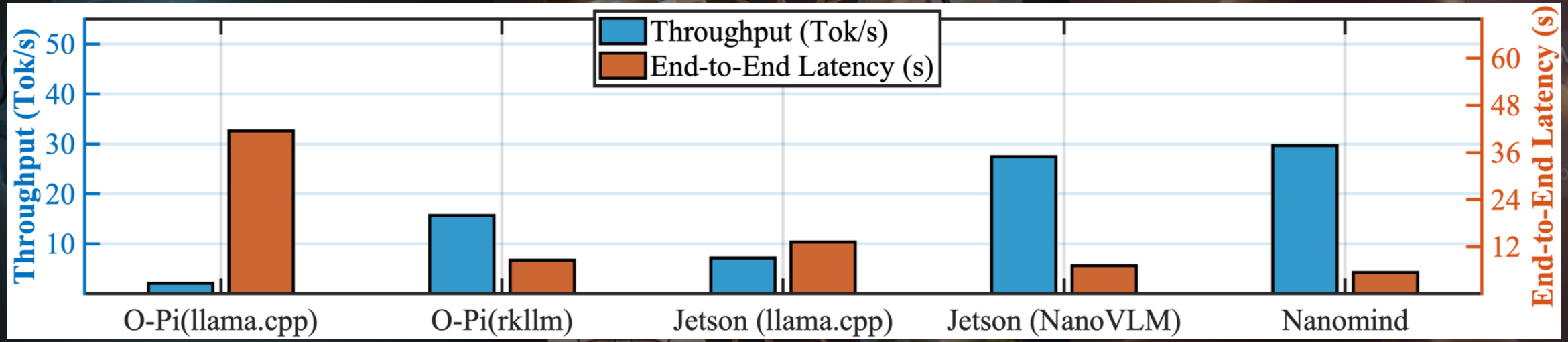


Prototype with Auto-focus Camera


Private Memory
Memoria: Self-Evolving Personal Agentic Memory
Text Conversation Chunks
Image Frames
Video Segments
Multimodal Input Stream
LLM Agent
Adaptive / Memoria Policy
Decide memory operations
CONTINUE
FLUSH
UPDATE
LINK
Decides memory operations, what to sit organize memory
RETRIEVAL PATH
Memory Recall Engine
Coarse-to-Fine Retrieval Idea
Episode Formation & Memory Consolidation
Episodic Memory
Memory Nodes
WRITE PATH
RETRIEVAL PATH
Event Nodes
TEMPORAL LINKS
Visual Evidence Store
Fine-grained Retrieval Bank
Retrieval Memory Bank
UPDATE
Personal Habits
Social Activity
Locations
Personal Memory
Memory System

Evaluation

Throughput (Tok/s)



Throughput (tokens/s) and end-to-end latency (s) for Qwen2-VL-2B-Instruct

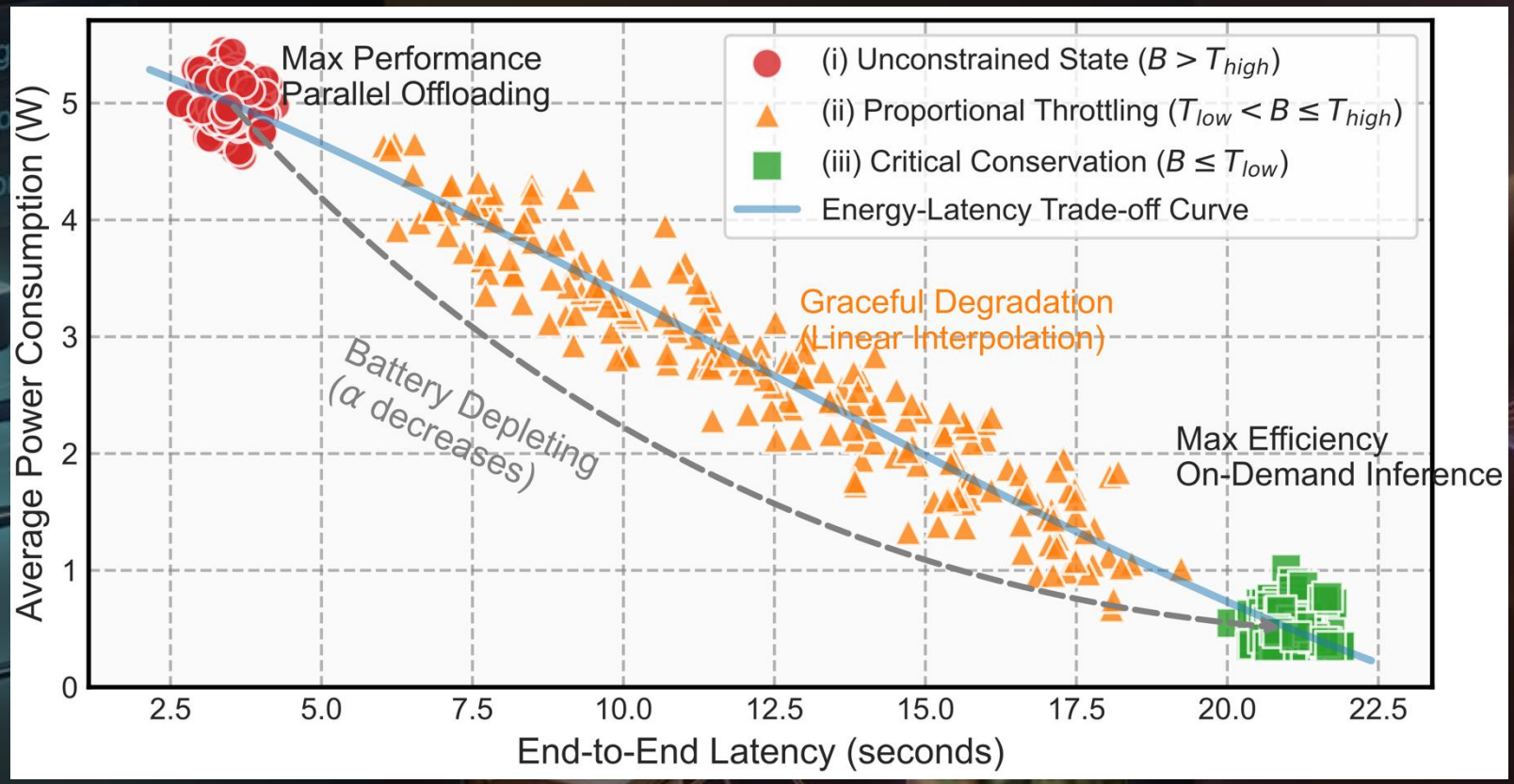
Efficient On-Device AI

Privacy-Preserving Personal Memory

Cognitive-State-Aware Assistance

Evaluation

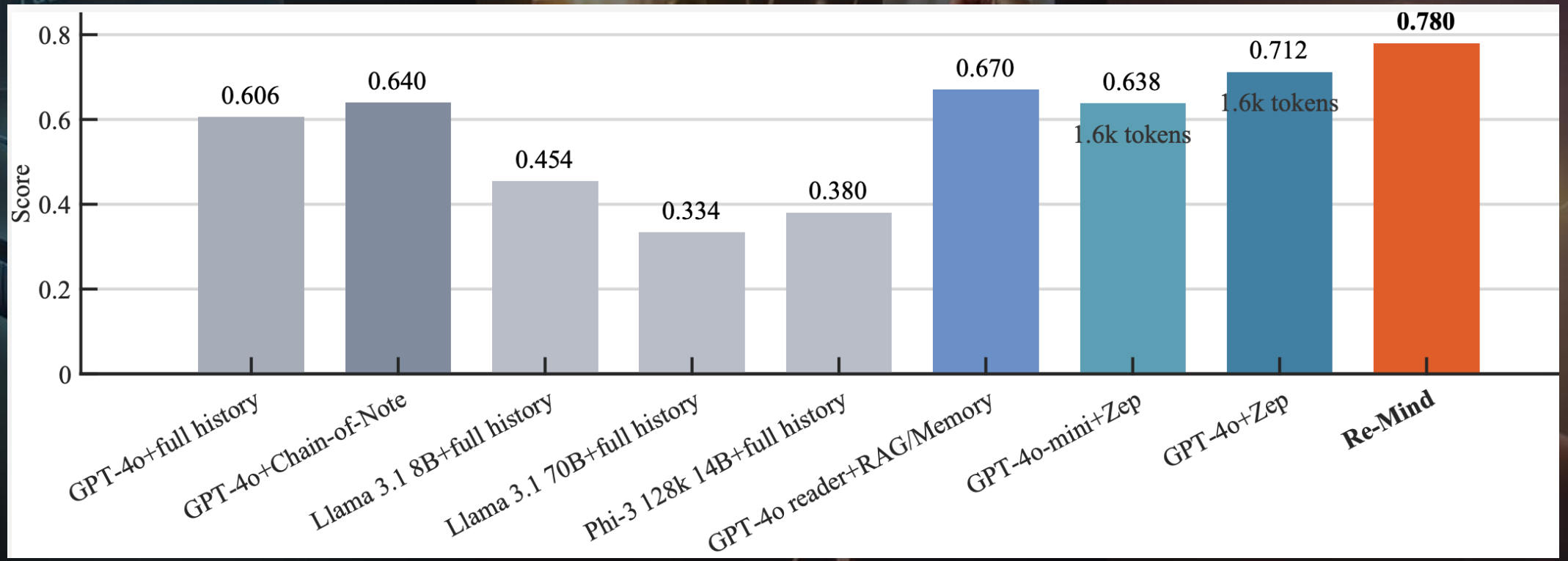
Energy-Latency Trade-off Across Different Modes



Energy–Latency Trade-off Across Three Power Modes. The curve illustrates how the system adapts to the battery level (B). (1) In the Unconstrained State, parallel acceleration delivers low latency at higher power. (2) In the Proportional State, th e system linearly throttles frame rate and memory bandwidth as B decreases, producing a continuous latency–power trade-off trajectory. (3) In the Critical State, the system transitions to the low-power On-Demand Cascade pipeline.

Evaluation

Evaluation of Long-Term Memory



LongMemEval-S long-context Evaluation Scores

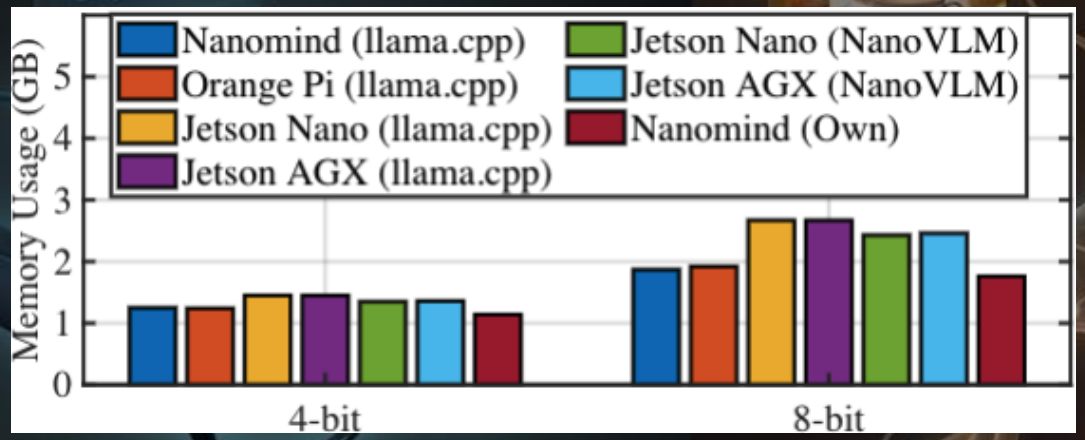
Efficient On-Device AI

Privacy-Preserving Personal Memory

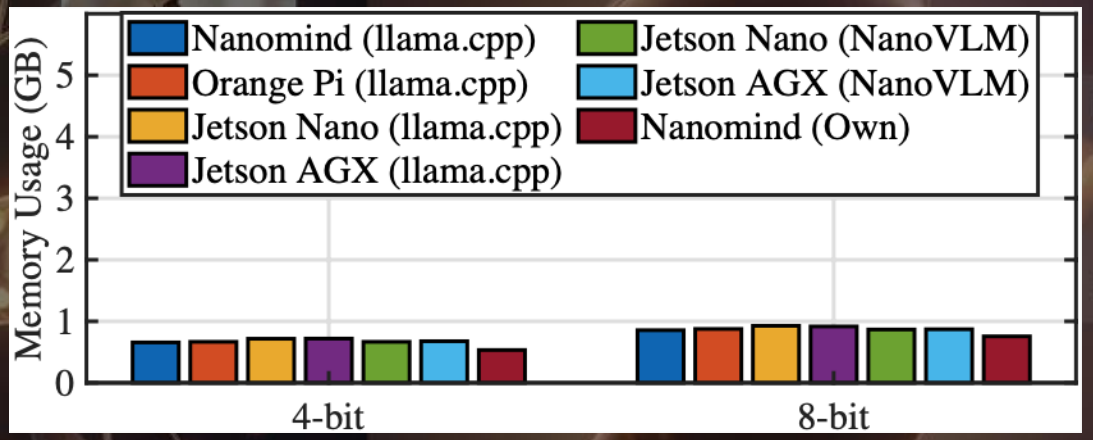
Cognitive-State-Aware Assistance

Evaluation

Memory Usage



Memory Usage (GB) while running Qwen2-VL-2B



Memory Usage (GB) while running Llava-onevision-0.5B