

# From Sight to Insight: Visual Memory for Smarter Assistants

Xin Luna Dong, Meta Reality Labs

GenAIRecP@KDD, 8/2025

This talk does not represent the company's point of view



# **What Is An Ideal Personal Assistant?**

# Recommendation & Personalization

## OBSERVE

what you do in your life

## UNDERSTAND

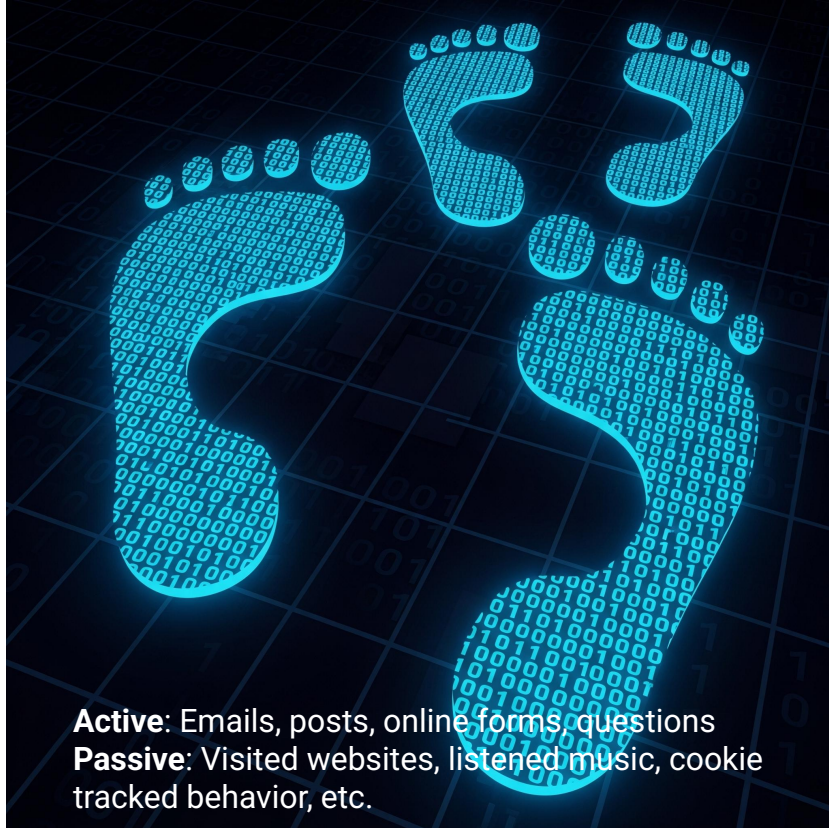
what you enjoy or find intriguing

## PROVIDE

what you need tailored to your interest



# Digital Footprints



**Active:** Emails, posts, online forms, questions

**Passive:** Visited websites, listened music, cookie tracked behavior, etc.

# Example: Social Network P13N



## OBSERVE

what you read and post in  
social network

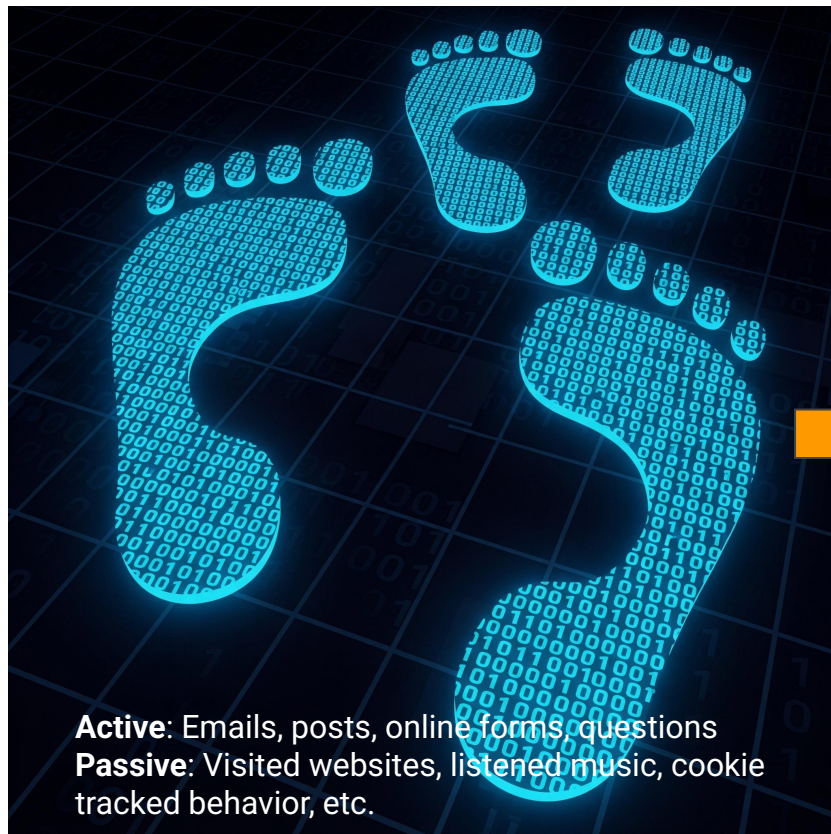
## UNDERSTAND

what you enjoy or find intriguing

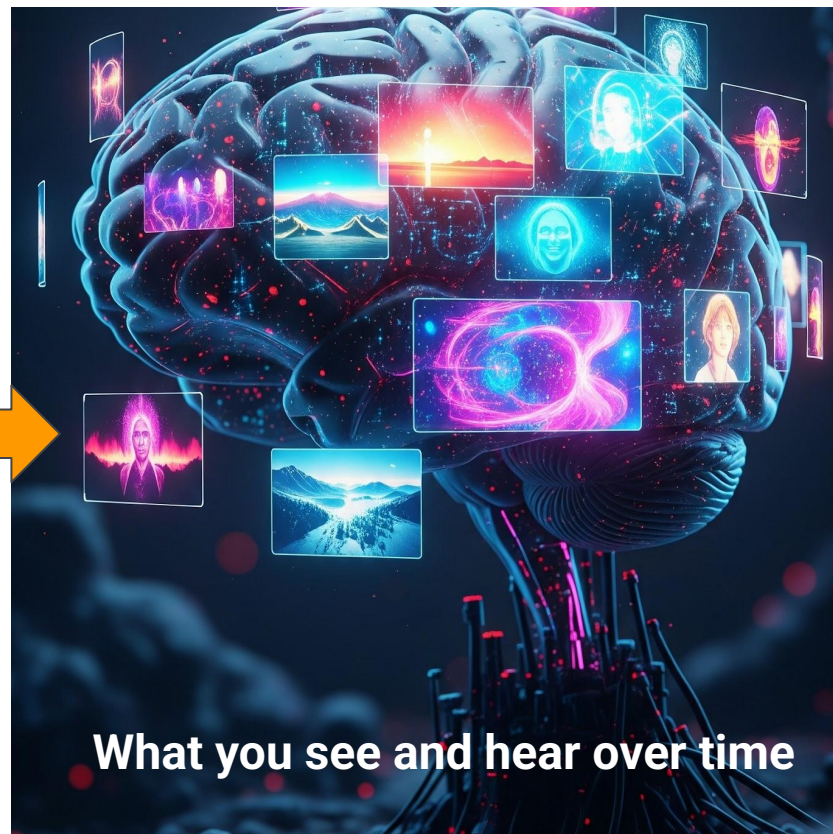
## PROVIDE

what you may enjoy reading

# Digital Footprints



# Visual Memory

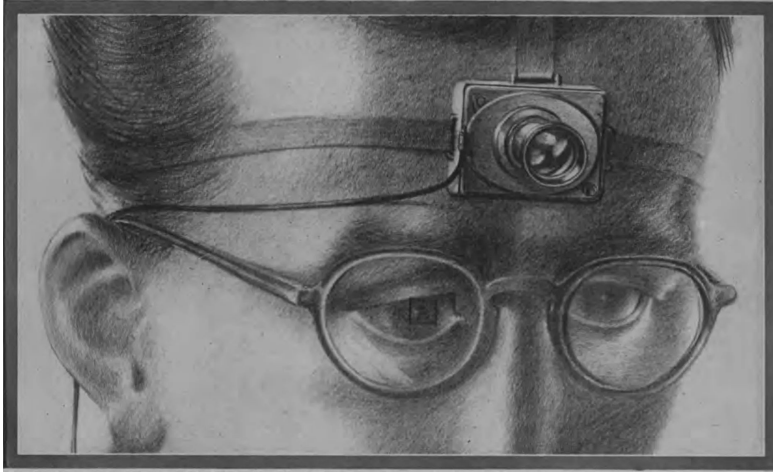




# What Can Visual Memory Provide In Addition?



# Visual Memory Is NOT IMPOSSIBLE



**MEMEX (MEMory & EXpansion)**  
*by Vannevar Bush (1945)*



**Wearables as great  
vehicles for life recording**



# Visual Memory QA & Personalization



## OBSERVE

what you see and hear

## UNDERSTAND

what you enjoy or find intriguing

## PROVIDE

personalized conversations

# Challenges in Recording & Leveraging Vis. Mem.

## OBSERVE

mostly on-device

- Hardware limitations
  - battery life
  - thermal constraints
  - storage capacity
  - transfer bandwidth

## UNDERSTAND

can be offline

- Diverse and noisy history
  - not relevant to a task
  - not accurately reflecting preferences
  - not preference-indicative or memory-worthy
- Special role of time and location

## PROVIDE

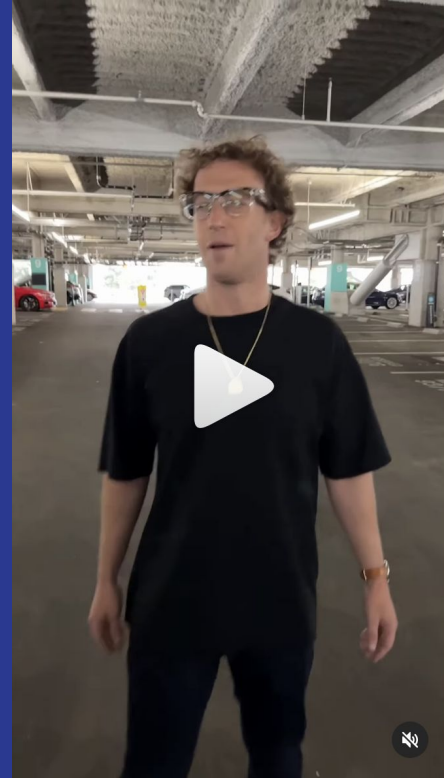
mostly on cloud

- Limited visual context windows for LLMs
- Runtime latency requirement
- Across-domain topics


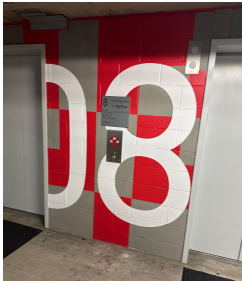


# Application I.

## Memory QA



# Memory QA Examples

Memory Question	Memory Retrieval Results	Answer
<i>What's the restaurant I saw last Thursday?</i>		The restaurant is named Mirate
<i>Which floor did I park?</i>		You parked on Floor 8.

# Memory QA—Remember This



## OBSERVE

what you ask to remember

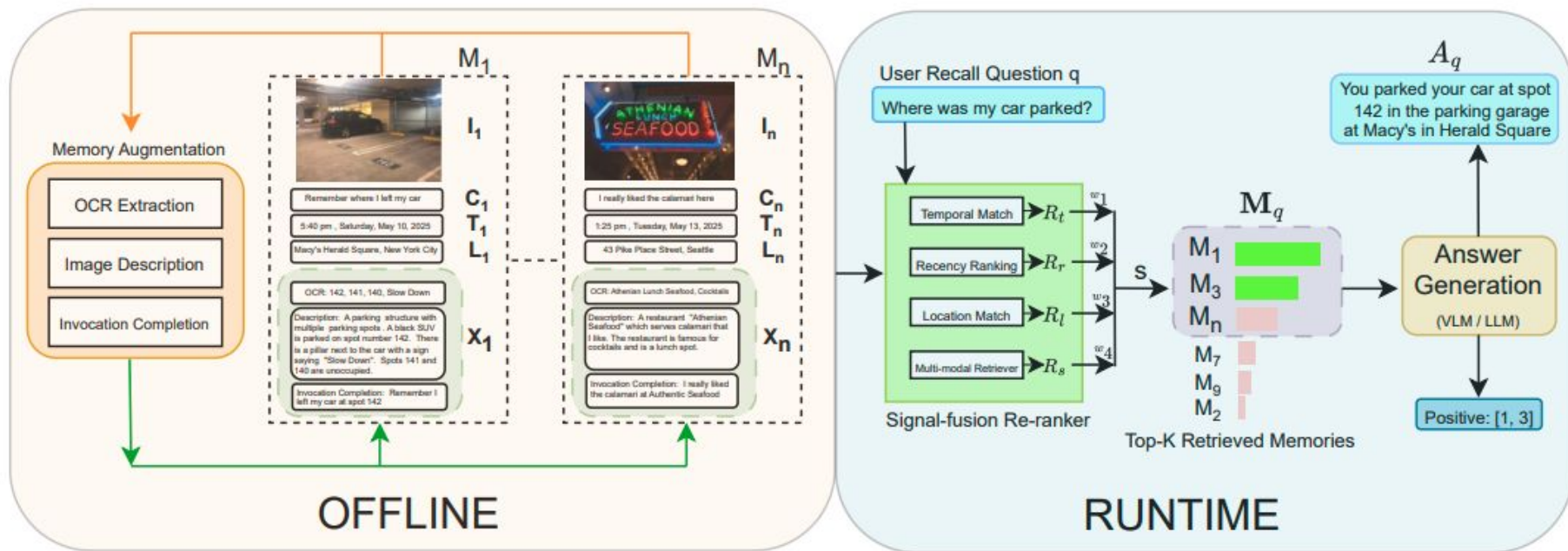
## UNDERSTAND

what is important to you

## PROVIDE

answers to your memory questions

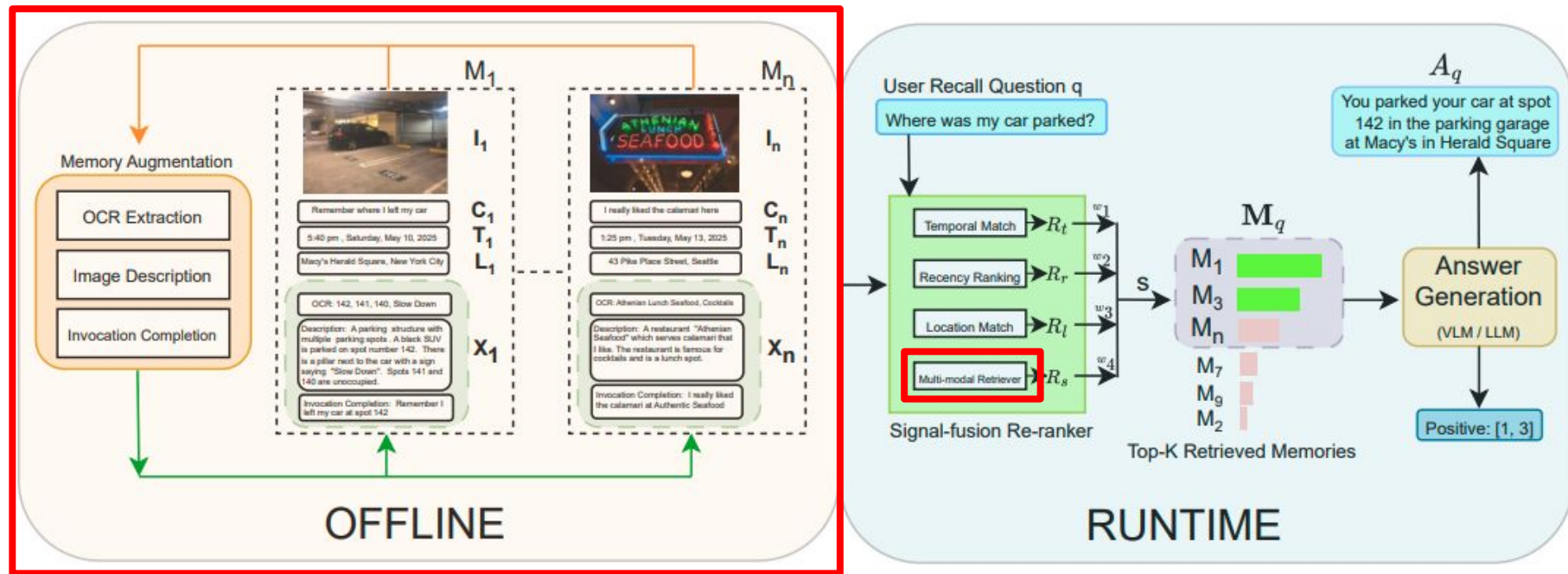
# The Pensieve Memory QA Solution



# RAG

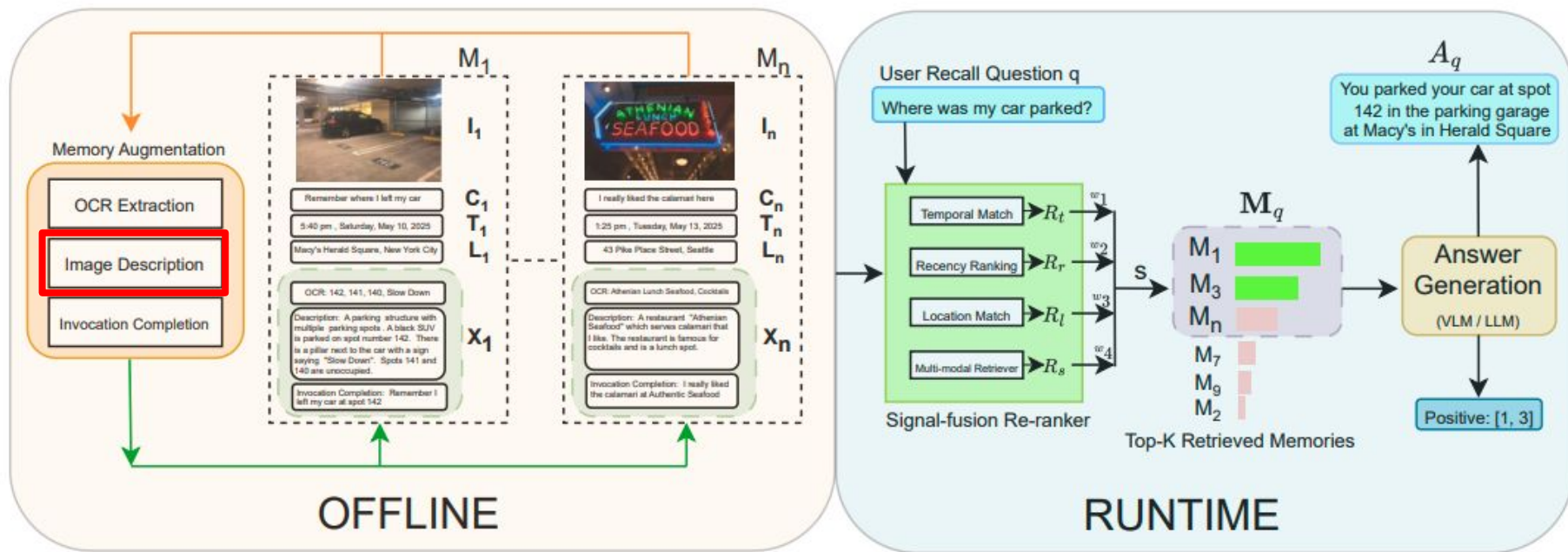


# The Pensieve Memory QA Solution



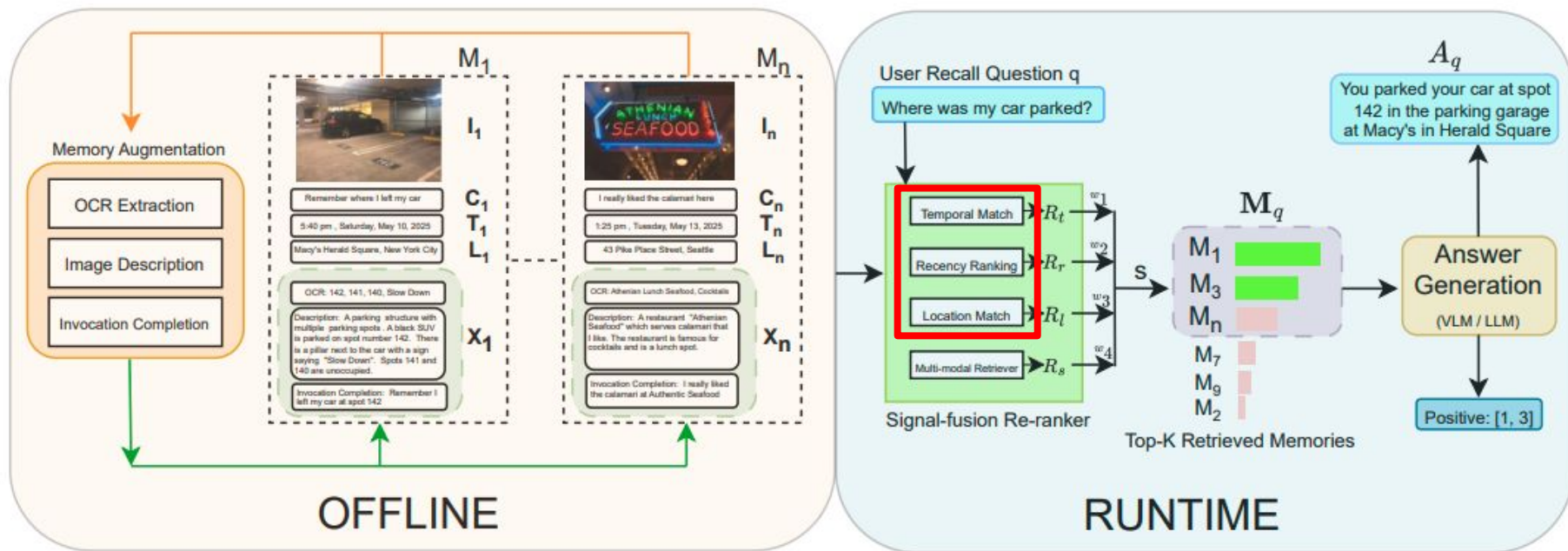
**Better memory selection** through offline augmentation & dual-modality retrieval

# The Pensieve Memory QA Solution



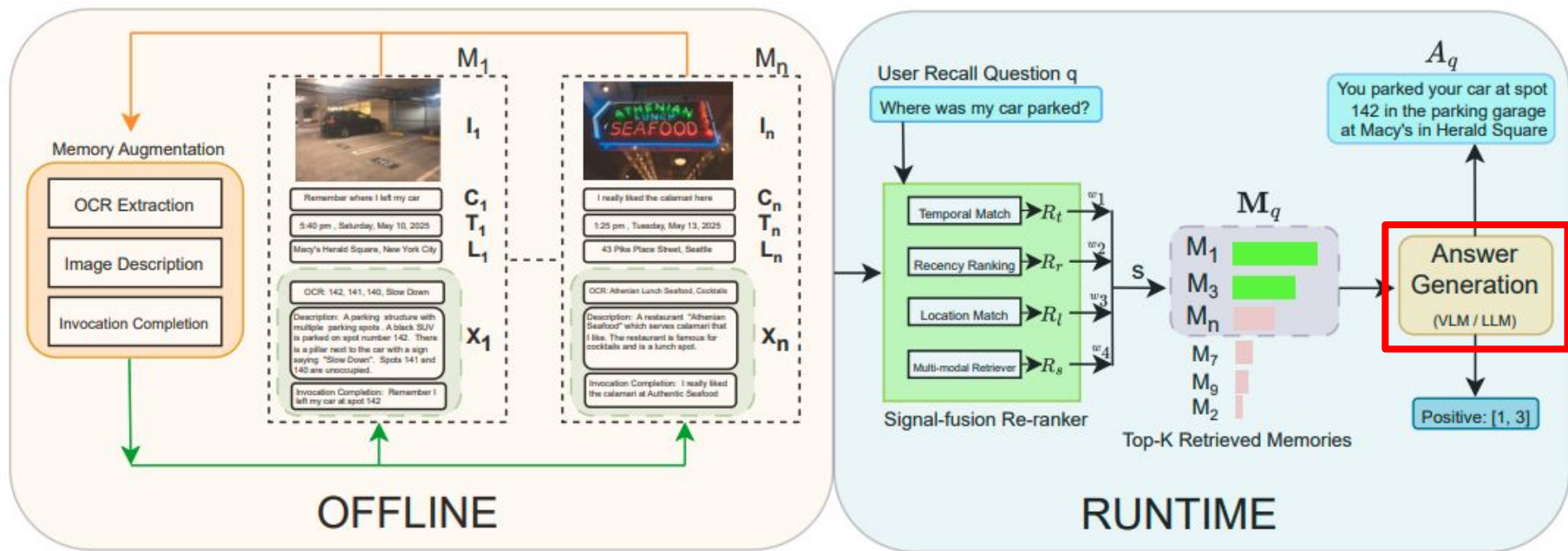
**More targeted to memory questions  
through recall object prediction**

# The Pensieve Memory QA Solution



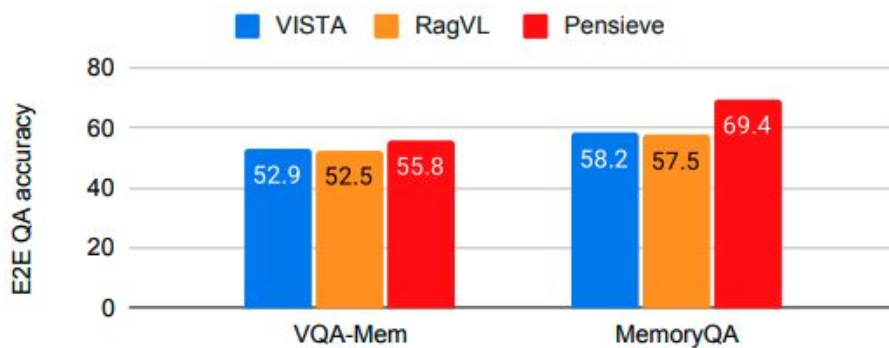
**Better understanding of temporal & location requirements**  
through temporal- & location-aware query rewriter and retrieval

# The Pensieve Memory QA Solution

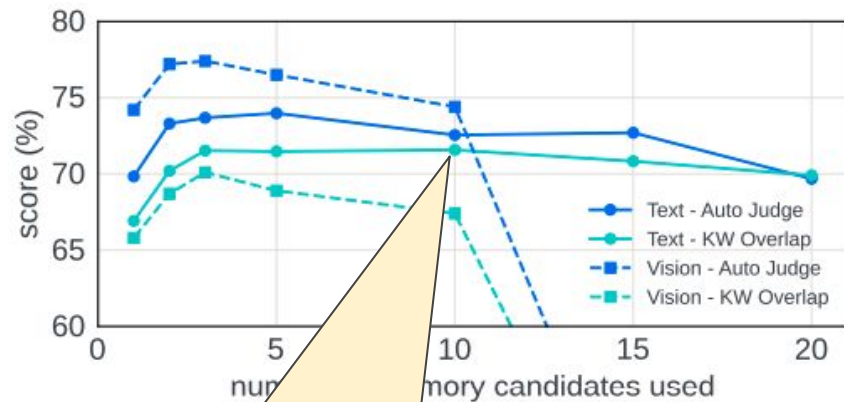


**Better answer generation and faster multi-memory aggregation through generation fine-tuning and text-based QA**

# The Pensieve Memory QA Quality



Improving over MM-RAG  
SOTA solutions by 11%



Text-based answer gen scales up  
to more candidate memories



# Memory QA—Pensieve Recap



## OBSERVE

*User-triggered*

## UNDERSTAND

*Augmentation by identifying memorization  
points and capturing details*

## PROVIDE

*User-triggered RAG w.  
multi-memory aggregation*



# Application II.

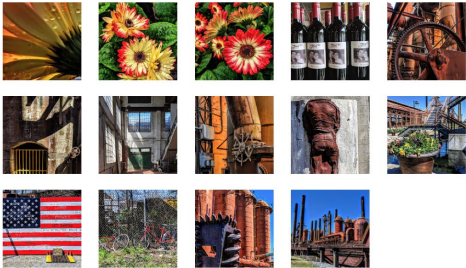
## P13N Through Visual History



Recommendation



# P13N Through Visual History Examples

Recommendation Question	Retrieval Candidates	Answer
<p>Which {museums, restaurants, parks, etc.} nearby should I visit?</p> <p>User visual history</p> 	<p>&lt;I1&gt; Vulcan Park and Museum, Statue symbolizing city's industries sits atop Red Mountain, surrounded by a park &amp; a museum.</p> <p>&lt;I2&gt; Sloss Furnaces Visitor Center, Blast furnace plant where iron was made from 1882–1971, now an arts &amp; education center with tours.</p> <p>&lt;I3&gt; Historic Cahaba Pumping Station, a working pump station located on the banks of the Cahaba River</p> <p>&lt;I4&gt; Alabama Jazz Hall of Fame, with a mission to foster, encourage, educate, and cultivate a general appreciation of the medium of jazz music</p>	<p>You may enjoy the Cahaba Pumping Station, an active facility on the riverbank showcasing local waterworks.</p> <p>Also consider Vulcan Park and Museum for industrial history, and Sloss Furnaces Visitor Center for a look at a preserved blast furnace.</p>

# VisualLens—Personalized Recomm.



## OBSERVE

which photos you take

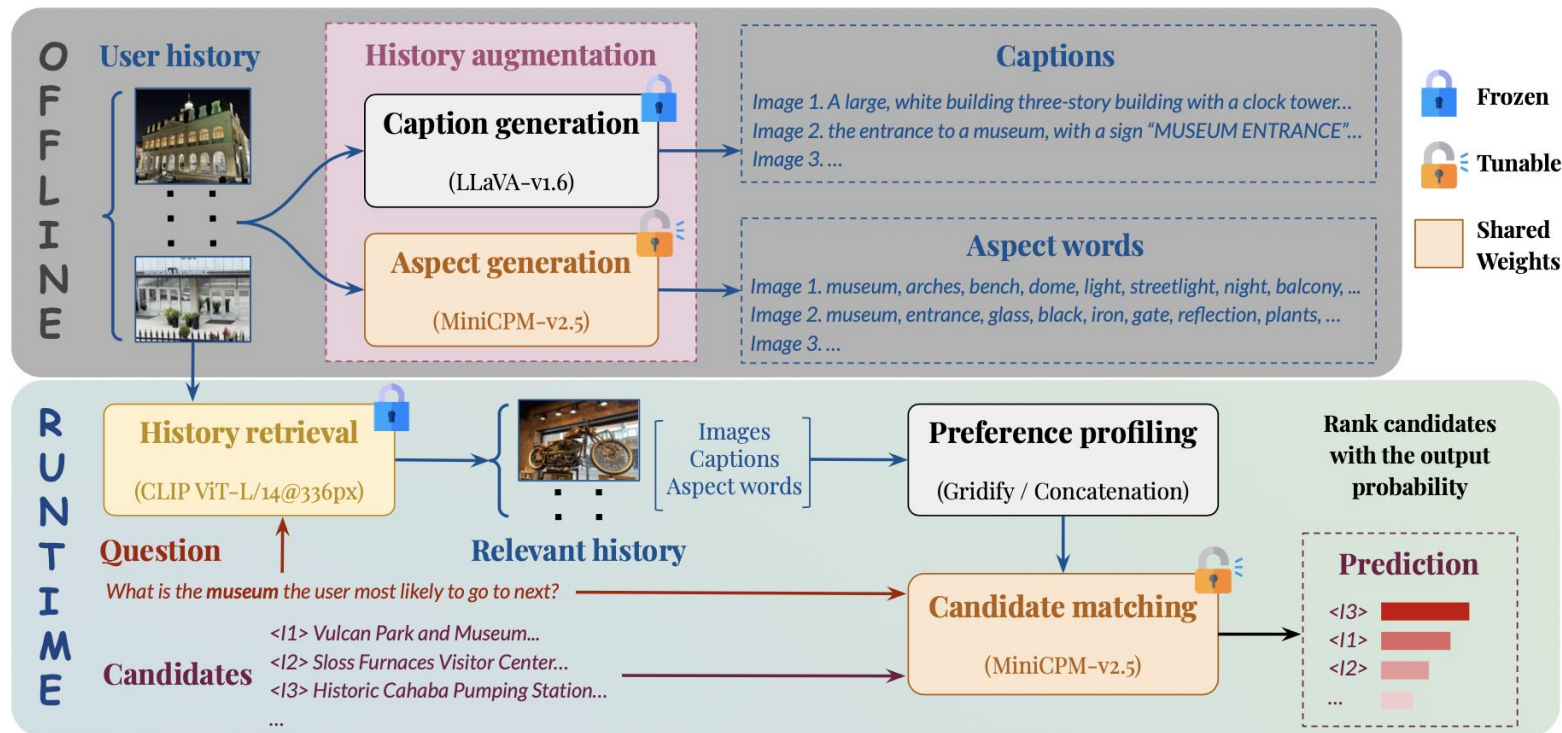
## UNDERSTAND

what you enjoy and find intriguing

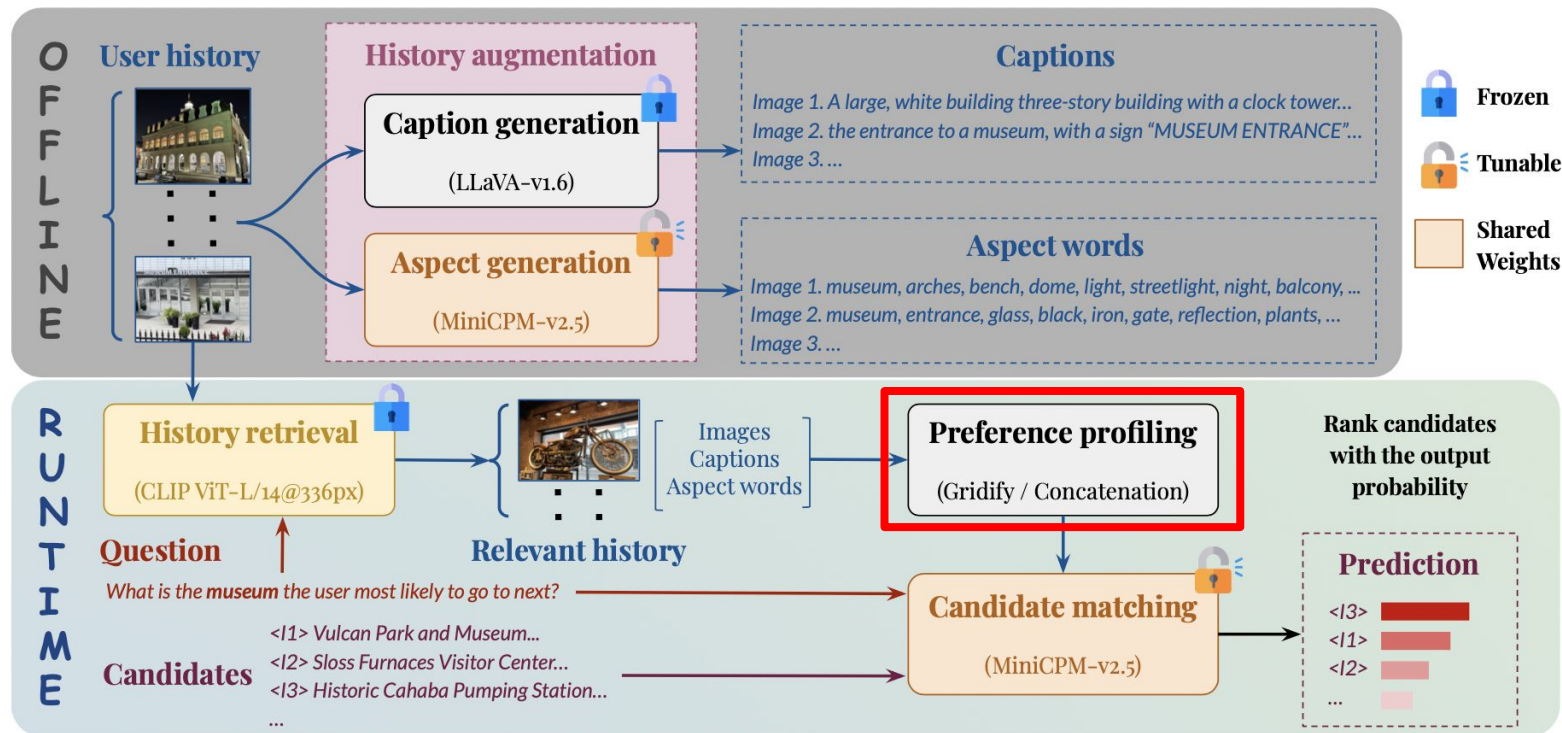
## PROVIDE

recommendations tailored to  
your interest

# VisualLens Method Overview

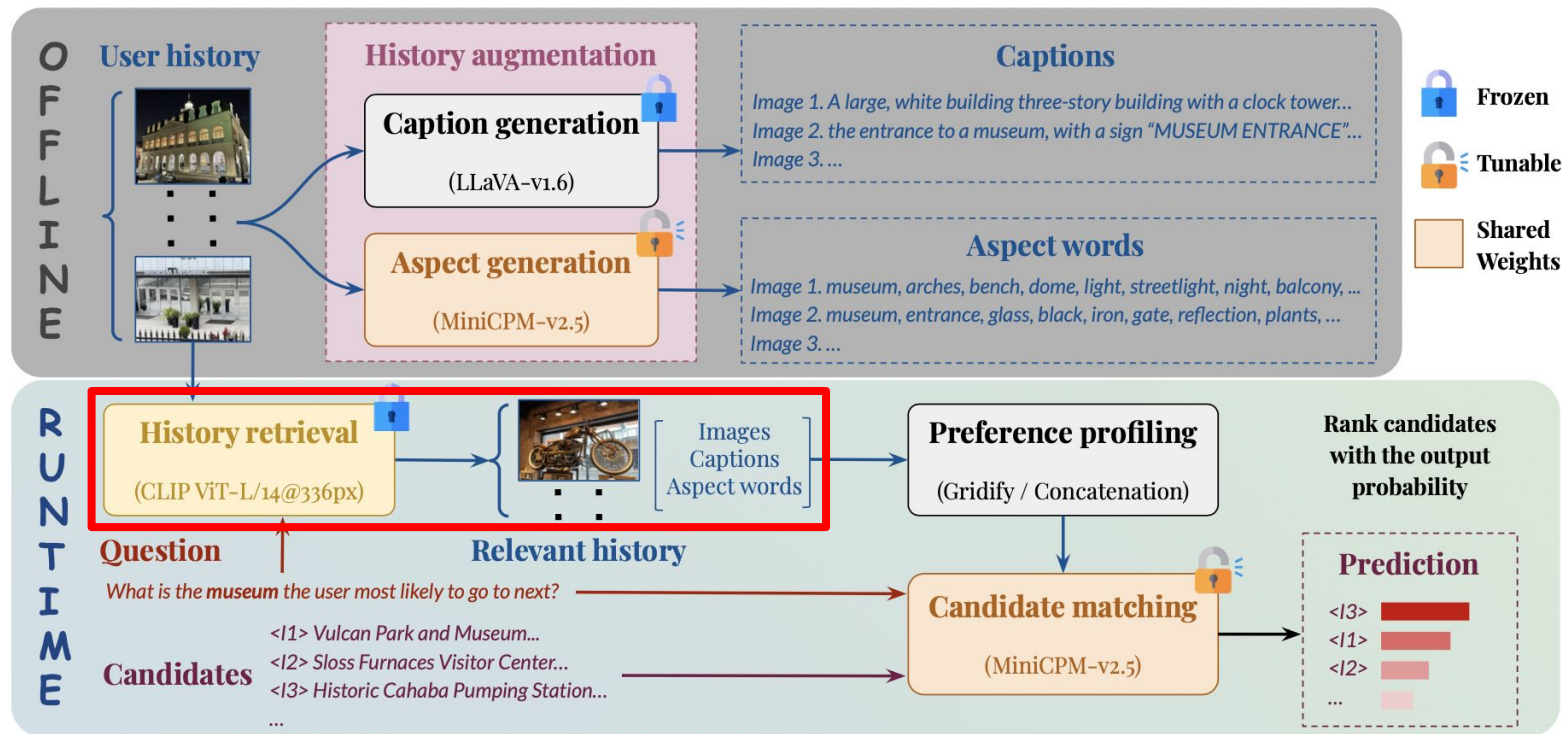


# VisualLens Method Overview



Efficient generation through  
grid-based multi-image inference

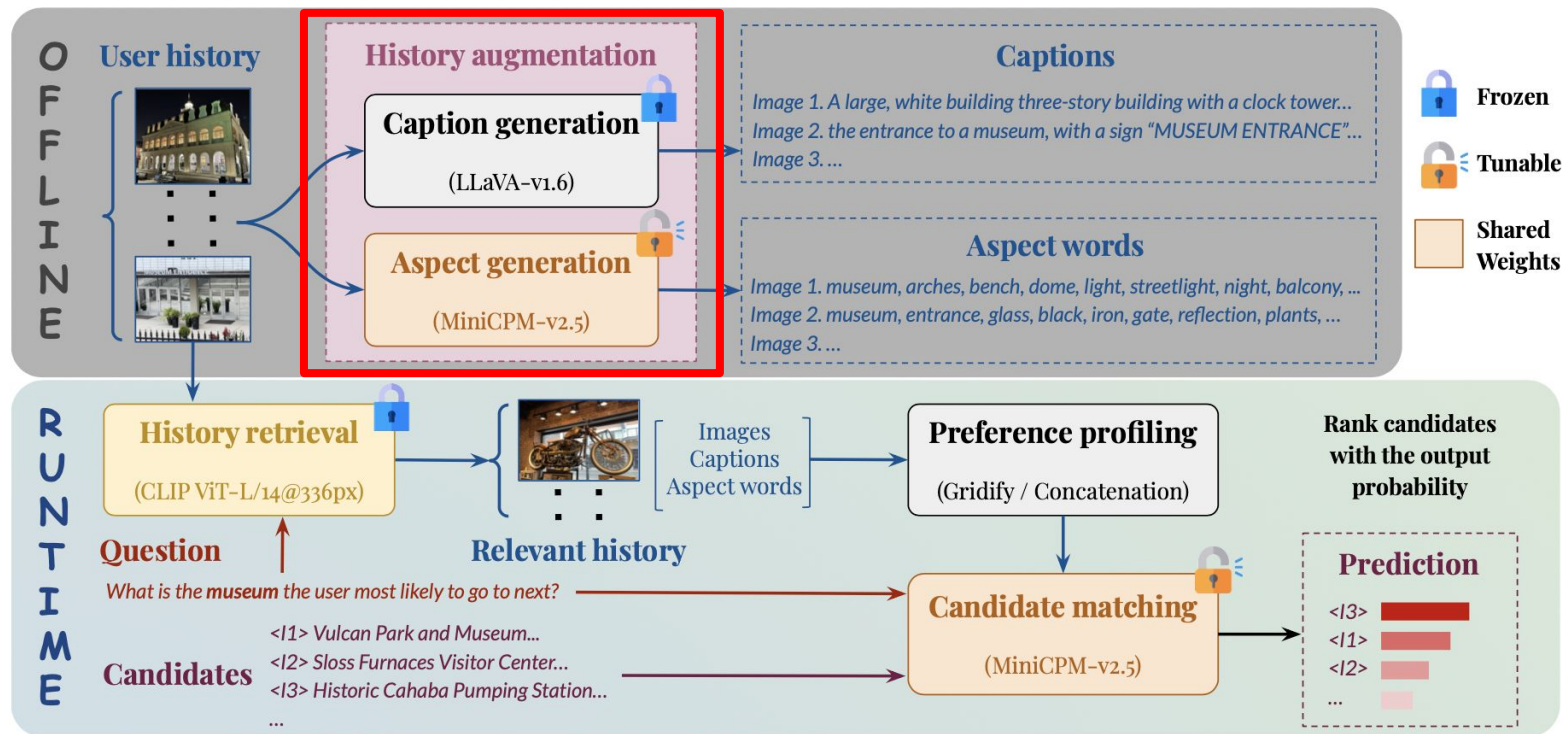
# VisualLens Method Overview



Relevant memories through history retrieval

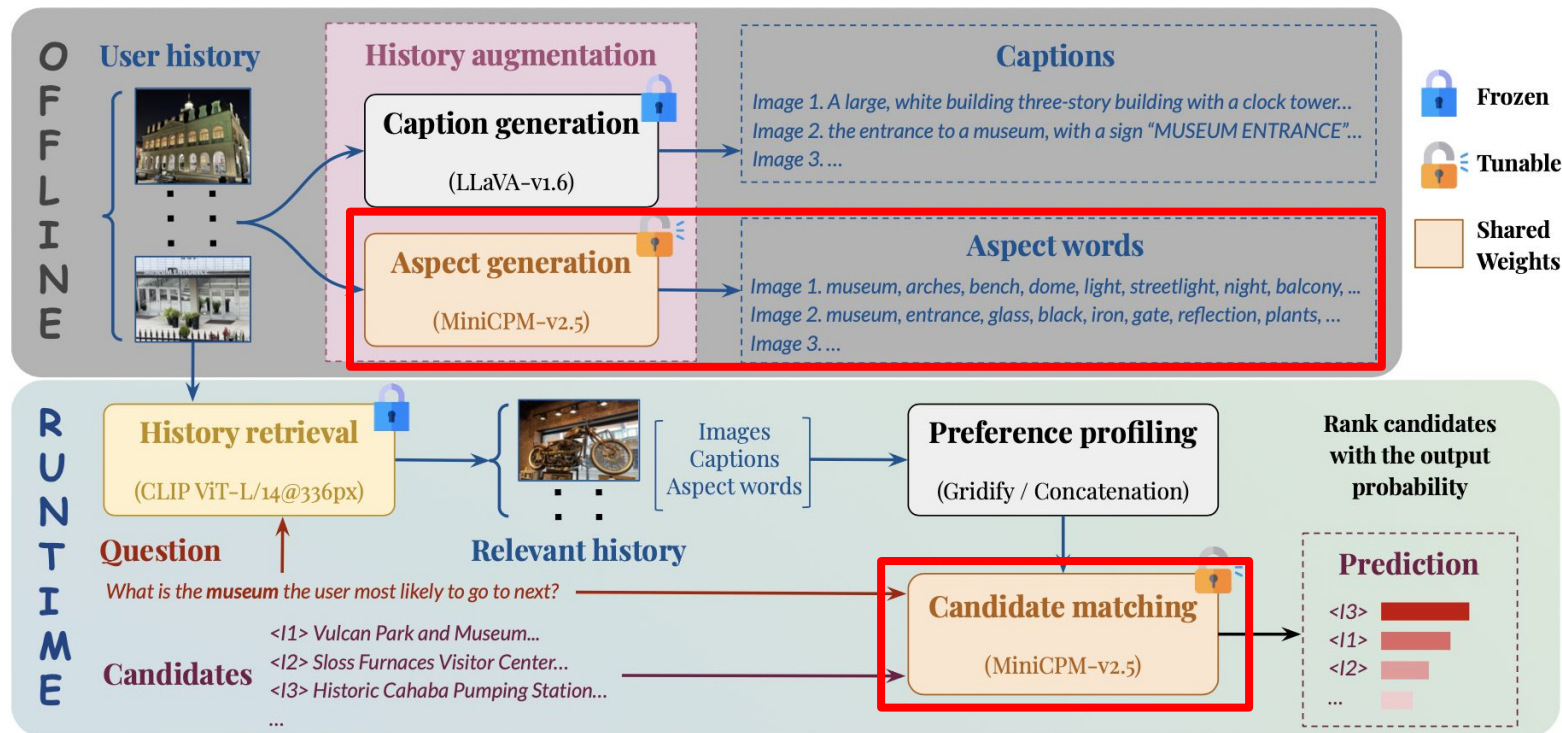


# VisualLens Method Overview



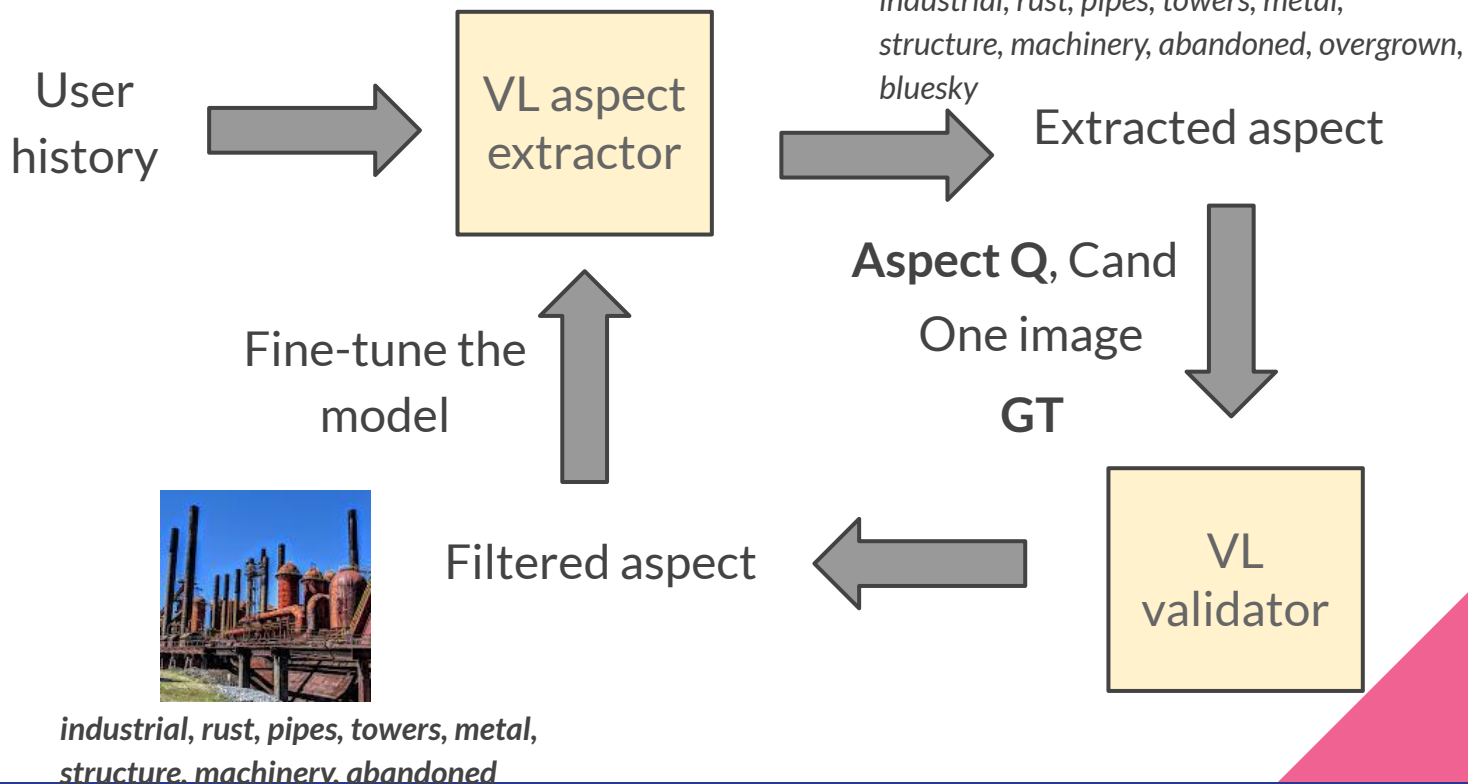
Better signal capturing through offline augmentation

# VisualLens Method Overview



**Better interest reflection** through joint fine-tuning of aspect word extraction and recommendation

# Aspect Self-Converge (Pos)



Aspect Q1: What are useful aspects to predict the **museums** the user will go after?

GT: Sloss Furnaces Visitor Center

Pred: **industrial, rust, pipes, towers, metal, structure, machinery, abandoned**

Aspect Q2: What are useful aspects to predict the **restaurants** the user will go after?

GT: Mugshots Grill and Bar - Birmingham, AL

Pred: **industrial, metal, structure**

# Aspect Self-Converge (Pos)



*industrial, rust, pipes, towers, metal,  
structure, machinery, abandoned*

User  
history

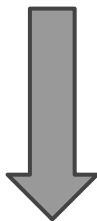


VL aspect  
extractor



Extracted aspect

Aspect Q, Cand  
One image  
GT

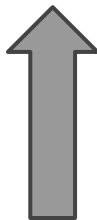


VL  
validator



Filtered aspect

Fine-tune the  
model



*industrial, rust, pipes, towers, metal,  
structure, machinery*

Aspect Q1: What are useful  
aspects to predict the  
**museums** the user will go  
after?

GT: Sloss Furnaces Visitor  
Center

Pred: **industrial, rust, pipes,  
towers, metal, structure,  
machinery**

Aspect Q2: What are useful  
aspects to predict the  
**restaurants** the user will go  
after?

GT: Mugshots Grill and Bar -  
Birmingham, AL

Pred: **industrial, metal,  
structure**

# Aspect Self-Converge (Pos)



industrial, rust, pipes, towers, metal,  
structure, machinery

User  
history



VL aspect  
extractor



Extracted aspect

Aspect Q, Cand  
One image  
GT

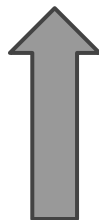


VL  
validator



Filtered aspect

Fine-tune the  
model



industrial, metal, structure,  
machinery

Aspect Q1: What are useful  
aspects to predict the  
**museums** the user will go  
after?

GT: Sloss Furnaces Visitor  
Center

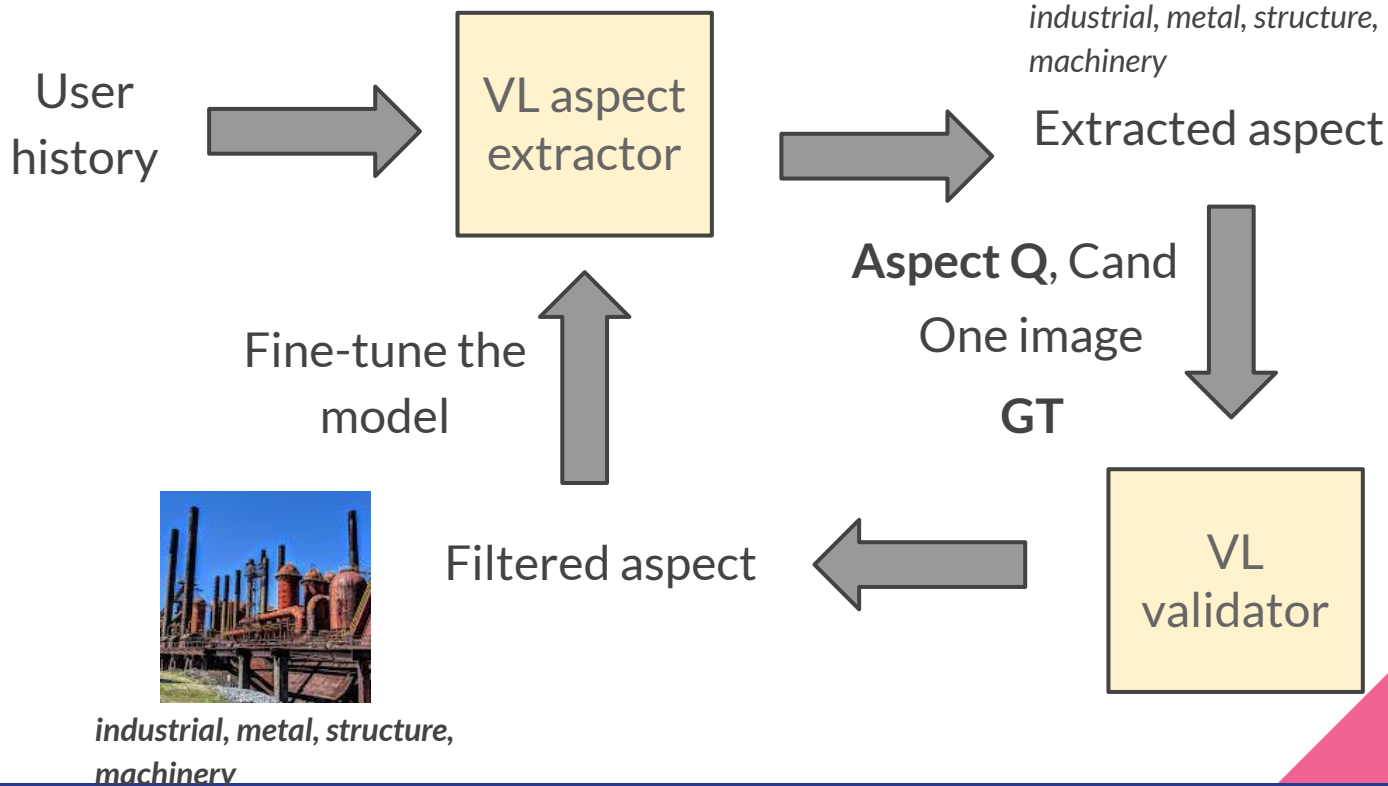
Pred: industrial, metal,  
structure, machinery

Aspect Q2: What are useful  
aspects to predict the  
**restaurants** the user will go  
after?

GT: Mugshots Grill and Bar -  
Birmingham, AL

Pred: industrial, metal,  
structure

# Aspect Self-Converge (Pos)



*industrial, metal, structure, machinery*

Aspect Q1: What are useful aspects to predict the **museums** the user will go after?

GT: Sloss Furnaces Visitor Center

Pred: *industrial, metal, structure, machinery*

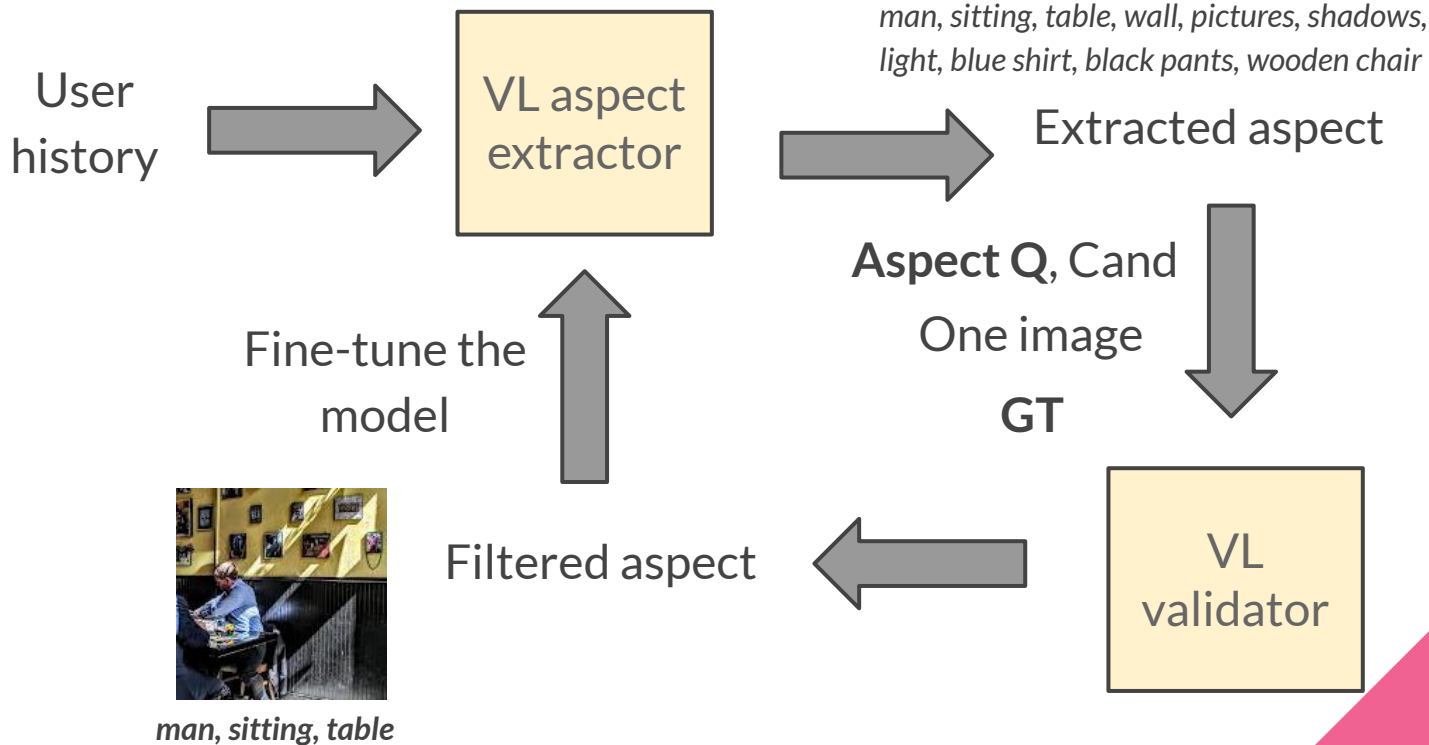
Aspect Q2: What are useful aspects to predict the **restaurants** the user will go after?

GT: Mugshots Grill and Bar - Birmingham, AL

Pred: *industrial, metal*



# Aspect Self-Converge (Neg)



Aspect Q1: What are useful aspects to predict the **museums** the user will go after?

GT: Sloss Furnaces Visitor Center

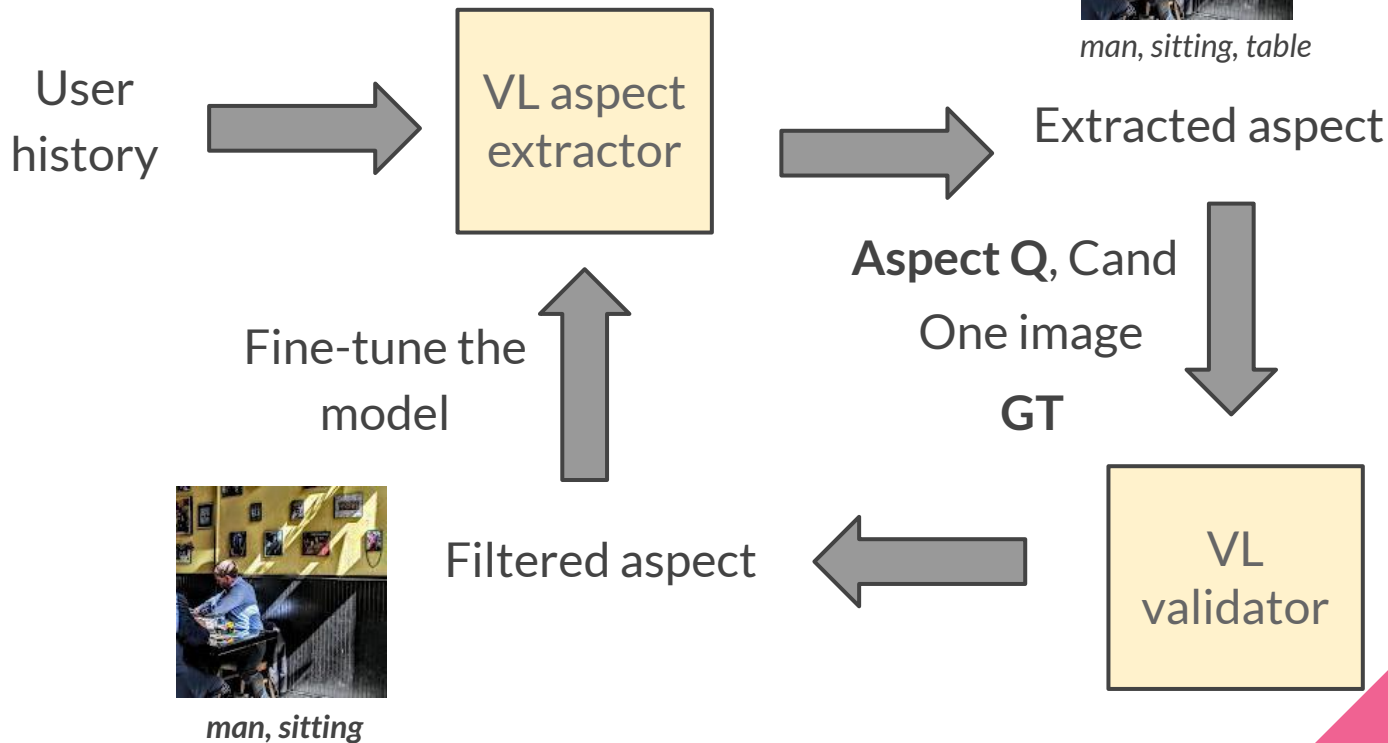
Pred: *man, sitting, table*

Aspect Q2: What are useful aspects to predict the **restaurants** the user will go after?

GT: Mugshots Grill and Bar - Birmingham, AL

Pred: *table, chair*

# Aspect Self-Converge (Neg)



Aspect Q1: What are useful aspects to predict the **museums** the user will go after?

GT: Sloss Furnaces Visitor Center

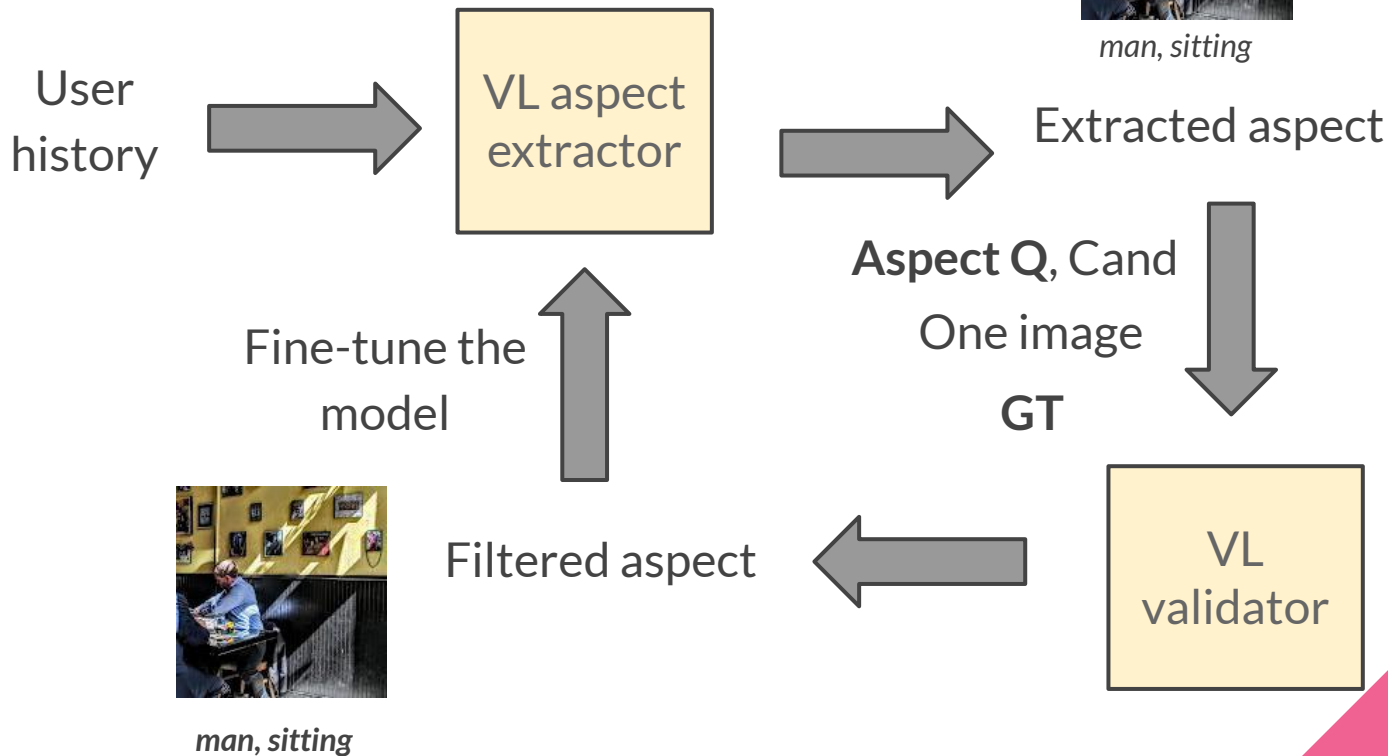
Pred: man,

Aspect Q2: What are useful aspects to predict the **restaurants** the user will go after?

GT: Mugshots Grill and Bar - Birmingham, AL

Pred: bar, sitting

# Aspect Self-Converge (Neg)



Aspect Q1: What are useful aspects to predict the **museums** the user will go after?

GT: Sloss Furnaces Visitor Center

Pred: **industrial, iron, statue, landmark, historic**

Aspect Q2: What are useful aspects to predict the **restaurants** the user will go after?

GT: Mugshots Grill and Bar - Birmingham, AL

Pred: **bar, burgers, beer**

# VisualLens Effectiveness

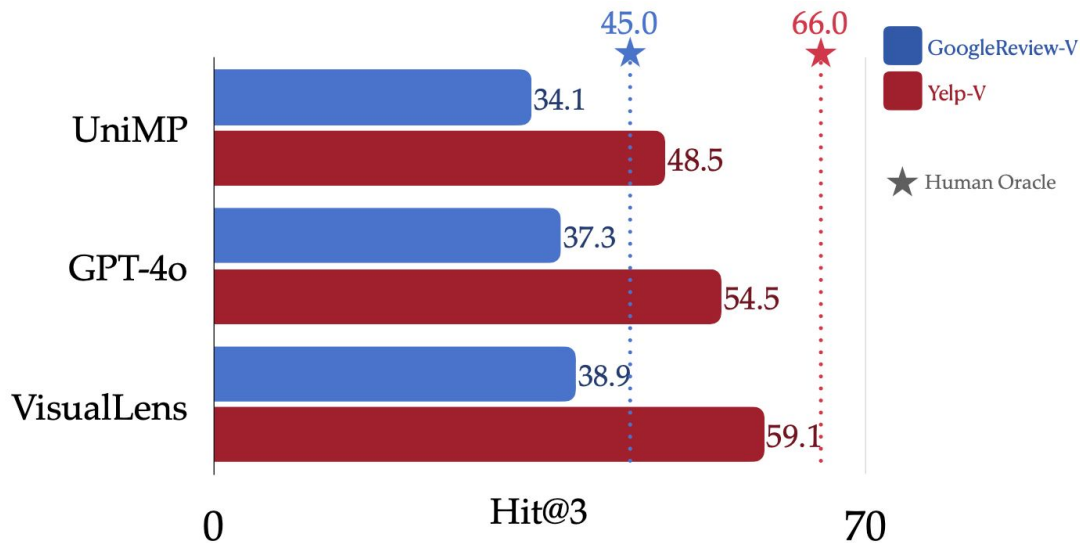
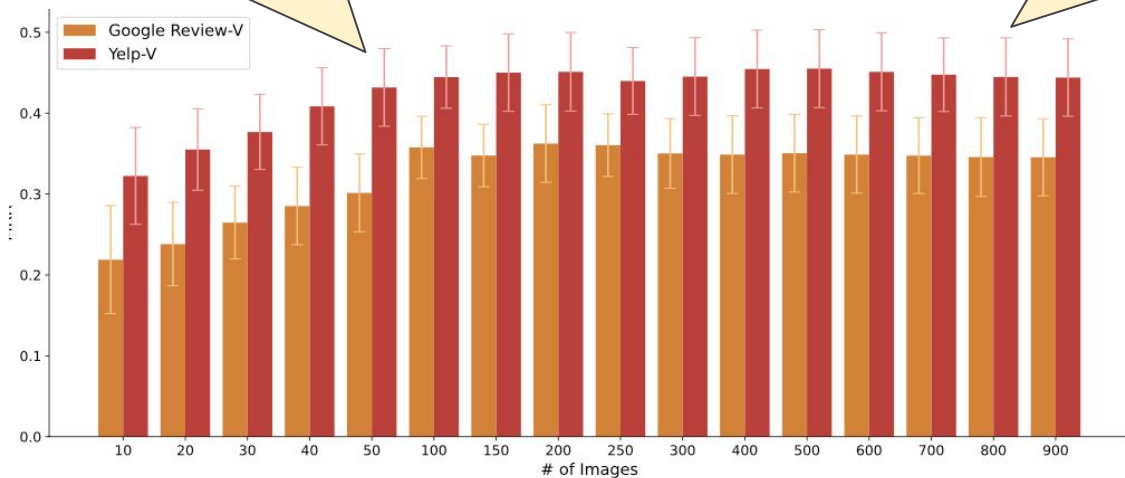


Figure 1. VISUALLENS leverages a user’s task-agnostic visual history to provide personalized recommendations. Our method outperforms GPT-4o by 1.6% ~ 4.6% on Hit@3.

# VisualLens Effectiveness vs. History Length

Quality improves with *richer* history

Robust against *longer* & *noisier* history



(b)

(b) MRR distribution over number of images. Both are on the User ID test set.

# Personalization—VisualLens Recap



## OBSERVE

*user-triggered,  
but not task-specific*

## UNDERSTAND

*augmentation by pinpointing  
user interest signals*

## PROVIDE

*Retrieve relevant memories as  
recommendation contexts*





# Proactive Memory Capture & Compression

# Always-on Proactive Capturing

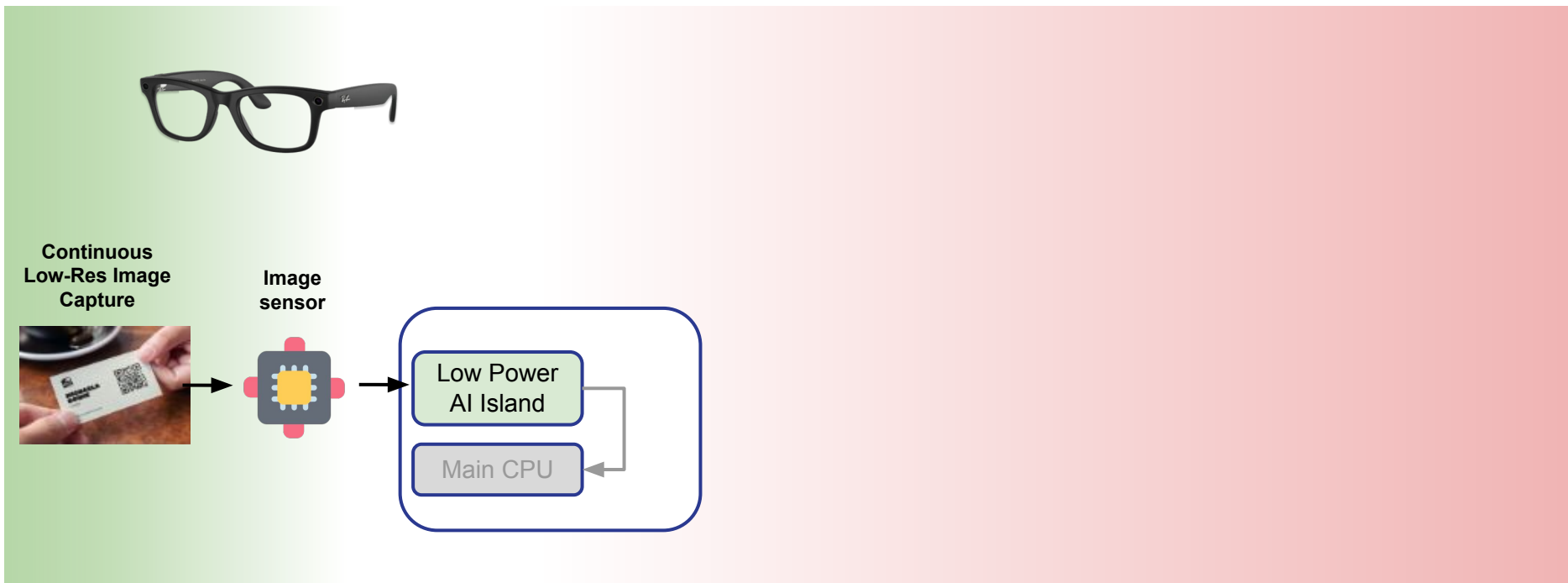
## Continuous video recording

- **Challenge:** Hardware limitations
  - battery life
  - thermal constraints
  - storage capacity
  - transfer bandwidth
- **Challenge:** Repeated info and needle in a haystack
- **Benefit:** Lossless information

## Periodic low-resolution photo capture

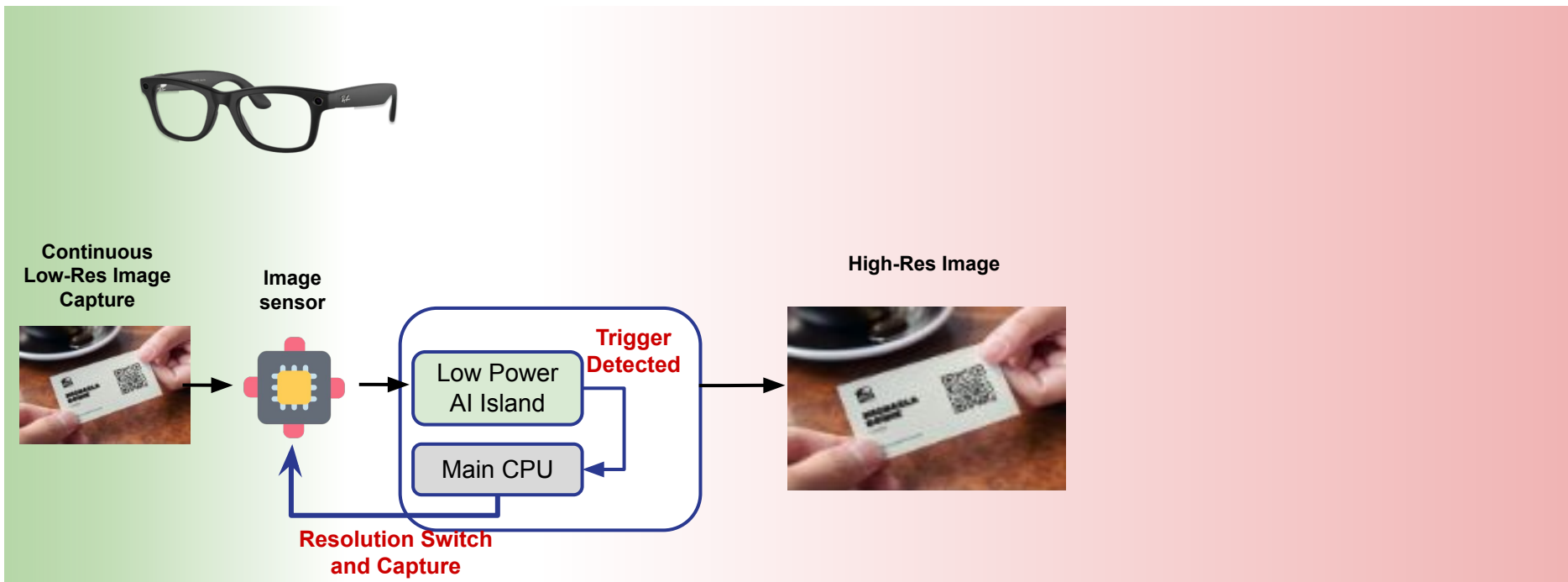
- **Challenge:** Lost information
  - frames
  - details
- **Challenge:** Still repeated info
- **Benefit:** Saved hardware resources

# Solution 1. Always-on Event-Triggered Proactive Capturing



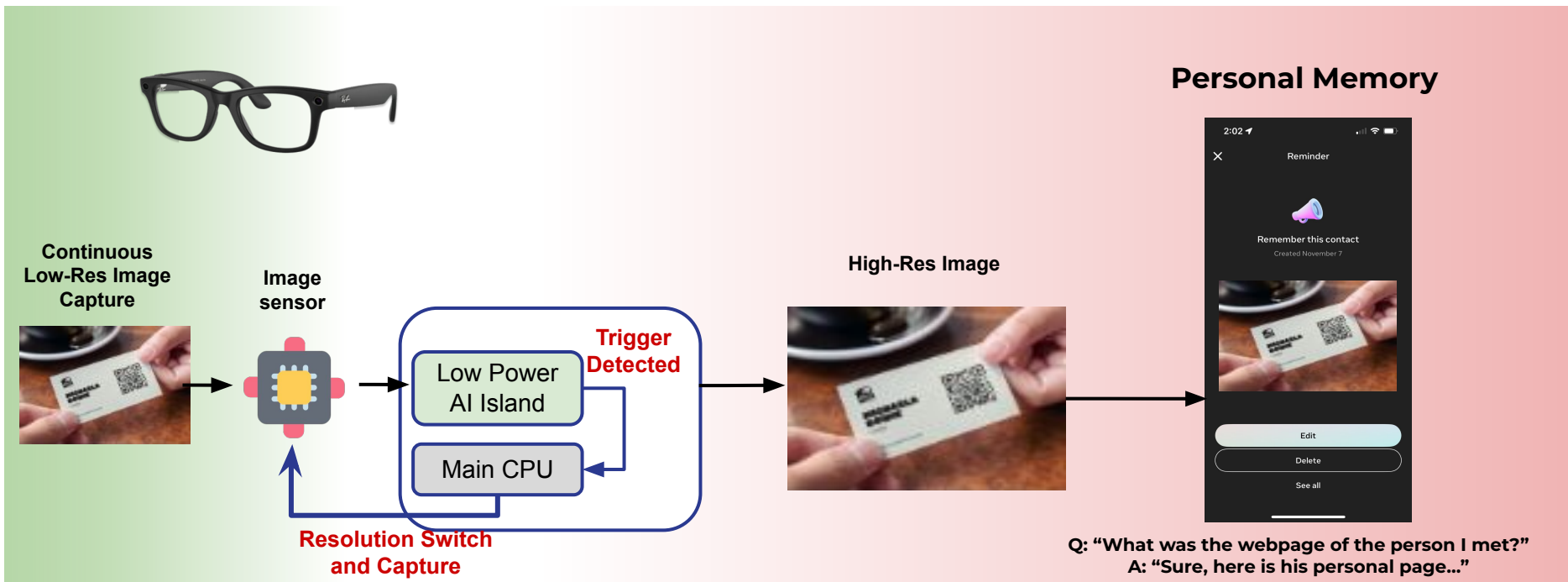
- Continuous low-resolution image capturing
- Event detection on low-power-AI-island

# Solution 1. Always-on Event-Triggered Proactive Capturing



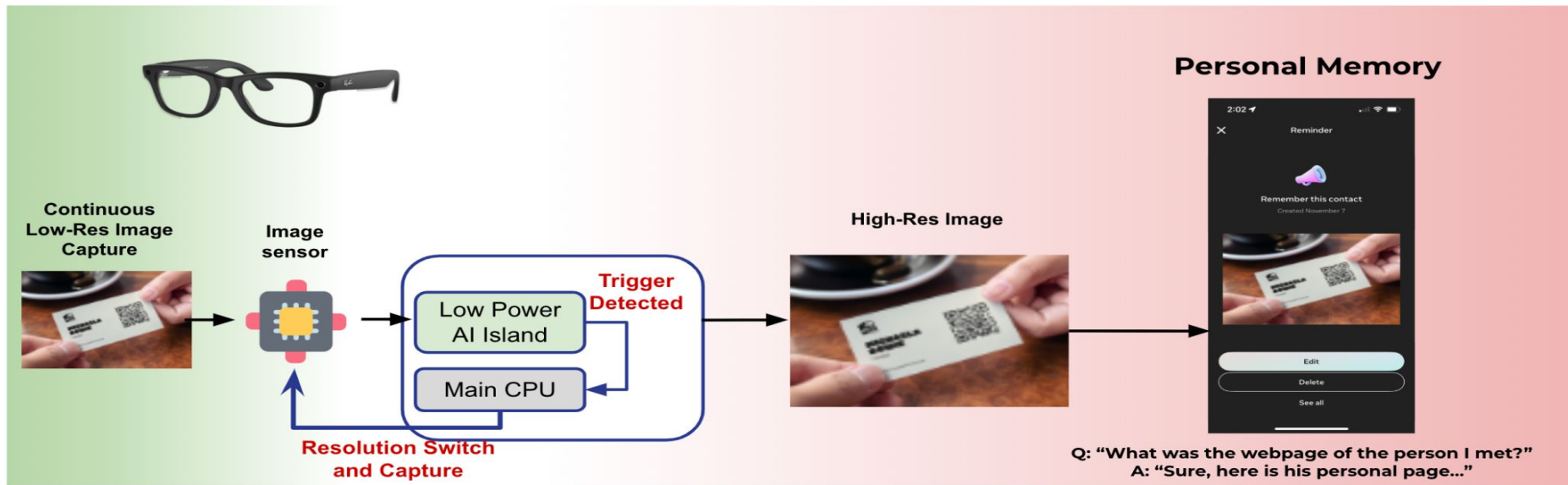
- Detected event wakes up main CPU
- CPU triggers high-resolution image capturing

# Solution 1. Always-on Event-Triggered Proactive Capturing



- High-resolution images transferred to personal memory
- Memory-QA on user's request

# Solution 1. Always-on Event-Triggered Proactive Capturing

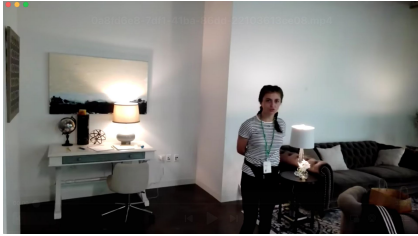


	Model Params	Model Size	Precision (HOI, AP50)	Precision* (OCR, AP50)	Precision (QR, AP50)
Low power island	1.78MB	1.88MB	~80%	~74%	~96%
Main CPU	14MB	-	~88%	~88%	-

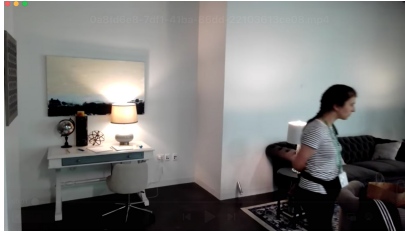
Memory Usage	2.9MB
Running Latency	74ms
Running Power	180mW



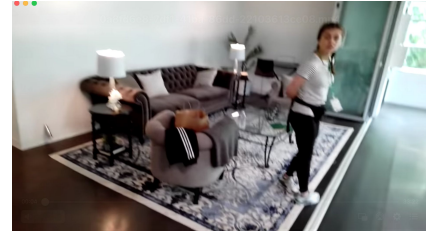
# Solution 2. Proactive Memory Compression



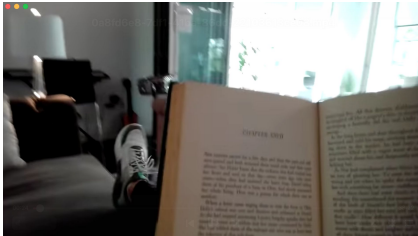
Frame1



Frame2



Frame3



Frame4

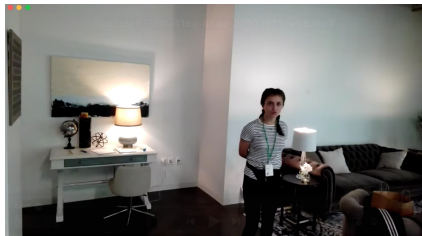


Frame5

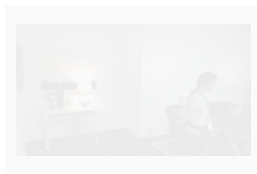


Frame6

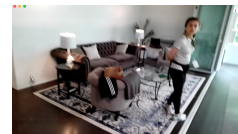
# Solution 2. Proactive Memory Compression



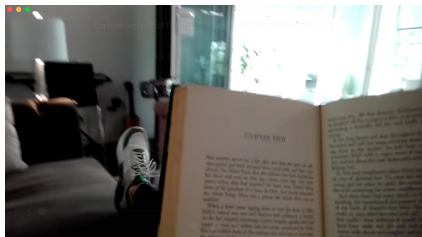
Frame1 - high



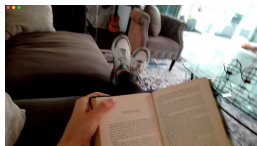
Frame2 - drop



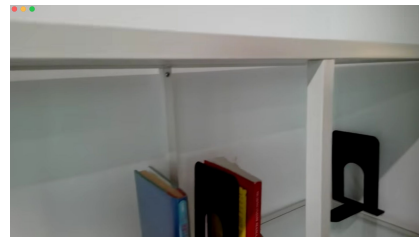
Frame3 - low



Frame4 - high



Frame5 - low



Frame6 - high

# Proactive Captured Memory



## OBSERVE

*always-on event-triggered  
proactive capturing*

## UNDERSTAND

*memory compression  
to reduce memory usage*

## PROVIDE

*memory answers  
and recommendations*



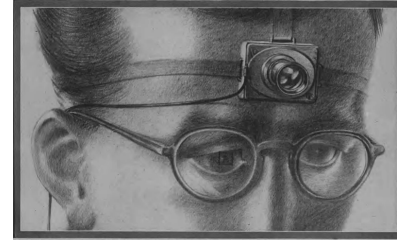
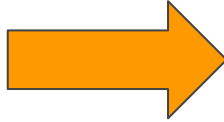
# **Towards Ultimate Visual Memory Enhanced Smart Assistant**

# Vision 1. Second Brain

Build a second brain to offload thoughts and knowledge for memory search, personalization, and memoir



**User-triggered data creation**



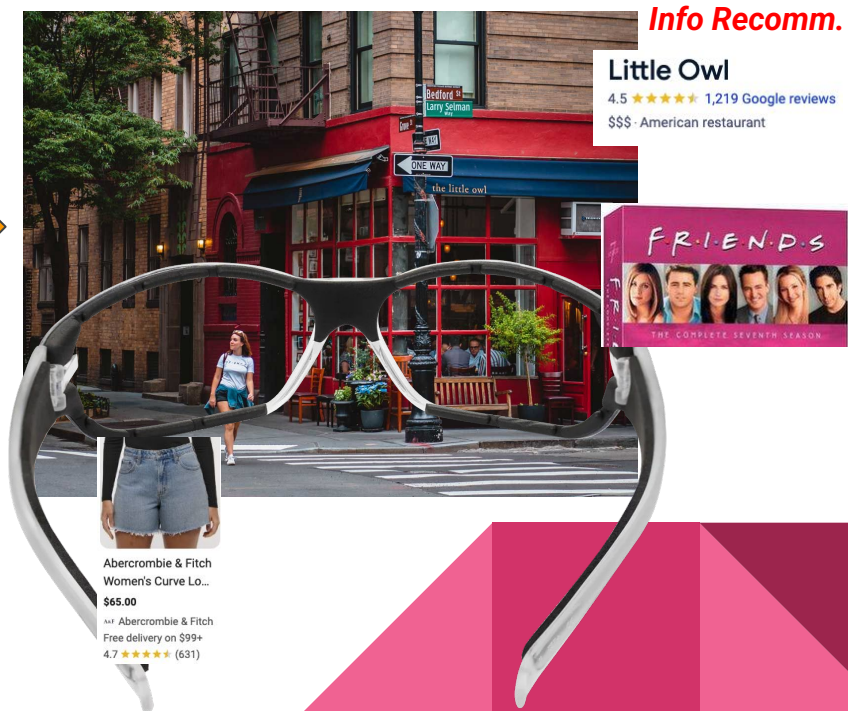
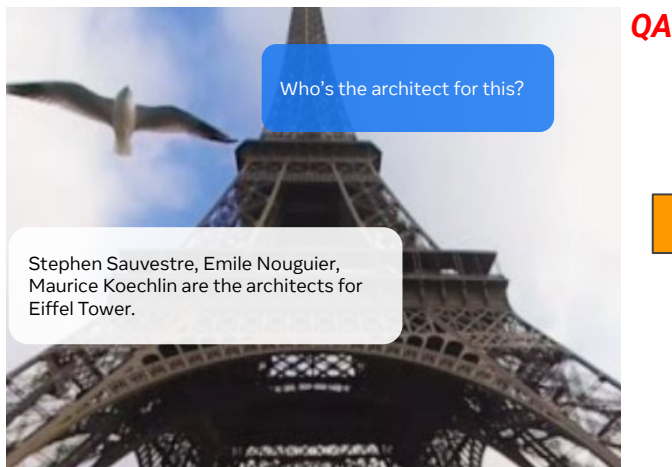
**Memex (1945)**



**Automatic data recording**  
**Digital footprint + Visual Memory**

# Vision 2. Smart Lens

Build smart lens to offer information relevant to what the user is seeing, saying, or asking, reactively and proactively





# Future Research Directions

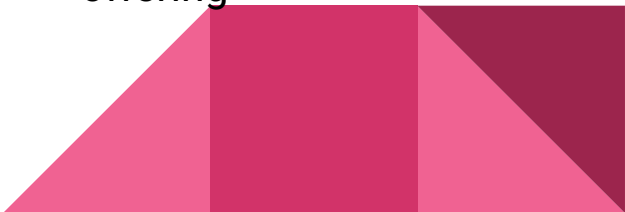
## OBSERVE

- Always-on event detection for proactive capturing
- Duplicate detection for capture and transmission suppression

## UNDERSTAND

- Memory compression to save space
- Memory organization and episode extraction for easier retrieval
- Seamless integration of multi-source memory
- Usage prediction for better augmentation
- Personality embedding for better recomm.

## PROVIDE

- Dense-memory search
  - Seamless blending of personal and public knowledge
  - Personalized LLM generation
  - Proactive information offering
- 

# Take-Aways

- Wearables devices provide a great vehicle for **life recording** and for **personalized assistance**
- **Visual memory** allows interesting memory QA (**Pensieve**) and personalized recommendation (**VisualLens**)
- We envision building **Second Brain** and **Smart Lens** based on **dense visual memory**

KDDCup Workshop—Comprehensive RAG Multi-modal Multi-turn Challenge. 8:00-12:00.

Towards a Knowledgeable Assistant: A Federated RAG Approach. MLoG, 9:55-10:30.

Managing Data, or Letting Data Manage Themselves. SKnowLLM, 3:00-3:30. Room: 717

# Thank You!