

# Next-Generation Intelligent Assistants for AR/VR Devices

Xin Luna Dong

8/2023

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

# Everyone Deserves An Assistant



# What is A Virtual Intelligent Assistant?

Respond to commands



“Hey Siri, set a timer to 7pm”

“Ok, added to today’s reminders”

# What is A Virtual Intelligent Assistant?

Control devices



“Hey Alexa, turn off bedroom lights”



# What is A Virtual Intelligent Assistant?

Provide information



“Hey, Google, when does summer end?”

“Summer ends on Thursday, Sept 22”

# Meta's Assistant

Empowering connection to people and experiences in your life

Meta Quest 2



“Hey Facebook” (double press the button on your controller)

“Who’s online?”--meet up with friends

“Open Beat Saber”--jump straight in the game, and more.

Ray-Ban Stories

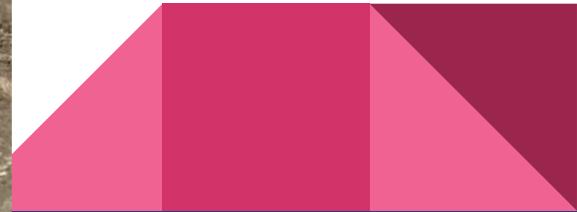


“Hey Facebook, take a picture” -- capture moments hands-free

“Hey Facebook”--call friends on Messenger, manage device settings, and more.

# Montage by Ray-Ban Stories

Empowering connection to people and experiences in your life



# GenAI Assistant

USER

I'm visiting long beach CA in august. give me a one-day itinerary

AI

Sure, I'd be happy to help you plan a one-day itinerary for your visit to Long Beach, CA in August!

Based on the information you provided, here's a sample itinerary that includes a mix of outdoor activities, cultural experiences, and delicious food:

9:00 AM - Start your day with a Segway tour of the shoreline, which you can rent from Long Beach Segway Tours. This is a fun and unique way to see the city, and you'll get to enjoy the ocean views and breezy weather.

10:30 AM - After your Segway tour, head over to the Naples canals and rent an electric Duffy Boat from London Boat Rentals. Pack a picnic lunch and enjoy a leisurely cruise around the canals, taking in the beautiful architecture and scenery.

12:30 PM - After your boat ride, head to Shoreline Village, a waterfront shopping and dining complex that offers a variety of restaurants and shops. Grab lunch at one of the many eateries, such as the Shoreline Village Cafe or the Long Beach Crab Shack.

2:00 PM - Spend the afternoon exploring the Long Beach Museum of Art, which is located nearby. The museum features a collection of American and European art, as well as rotating exhibitions.

4:00 PM - Take a break and enjoy a coffee or snack at the museum's cafe, or head back to Shoreline Village for some shopping or browsing.

USER

What should I present in the tutorial "Towards Next-Generation Intelligent Assistants leveraging LLM techniques"

AI

Based on the given web results, the tutorial "Towards Next-Generation Intelligent Assistants leveraging LLM techniques" could cover the following topics:

1. Introduction to LLM (Large Language Model) techniques and their applications in Intelligent Assistants.
2. Overview of the current state-of-the-art in Intelligent Assistants, including popular voice assistants like Alexa, Google Assistant, and Siri.
3. Discussion of the limitations and challenges of current Intelligent Assistants, such as their inability to understand complex requests, lack of common sense, and poor handling of multi-modal input.
4. Explanation of how LLM techniques can help address these challenges, including the use of hyperbolic graph neural networks, graph and geometry generative modeling, and topic modeling.
5. Presentation of case studies or research papers that demonstrate the effectiveness of LLM techniques in improving Intelligent Assistants, such as improved natural language understanding, more accurate response generation, and enhanced user experience.
6. Discussion of the ethical and societal implications of using LLM techniques in Intelligent Assistants, such as privacy concerns, bias, and the potential for job displacement.
7. Hands-on tutorial or demonstration of how to implement LLM techniques in Intelligent Assistants using popular frameworks and libraries like TensorFlow, PyTorch, or Gensim.
8. Future outlook and trends in LLM techniques for Intelligent Assistants, including the potential for multimodal interaction, integration with other AI technologies like computer vision and robotics, and the role of Explainable AI.
9. Open discussion and Q&A session to address audience questions and concerns.

By covering these topics, the tutorial "Towards Next-Generation Intelligent Assistants leveraging LLM techniques" can provide a comprehensive overview of the current state and future directions of Intelligent Assistants, as well as practical guidance on how to apply LLM



# What is An Ideal Virtual Intelligent Assistant?

An *intelligent assistant* should be an agent that **knows you and the world**, can **receive your requests** or **predict your needs**, and provide you the **right services at the right time** with your permission.



# Stages of Intelligent Assistant

## I. Chatbot

Text input



## II. Voice Asst

Voice input



## III. AR/VR Asst

Voice + Visual + Context

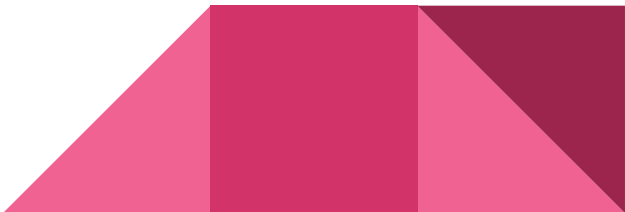


# Structure of the Talk

## Outline

- What is an Intelligent Assistant?
- Techniques to support current intelligent assistants
- Challenges and initial solutions for the next generation of multi-modal assistants

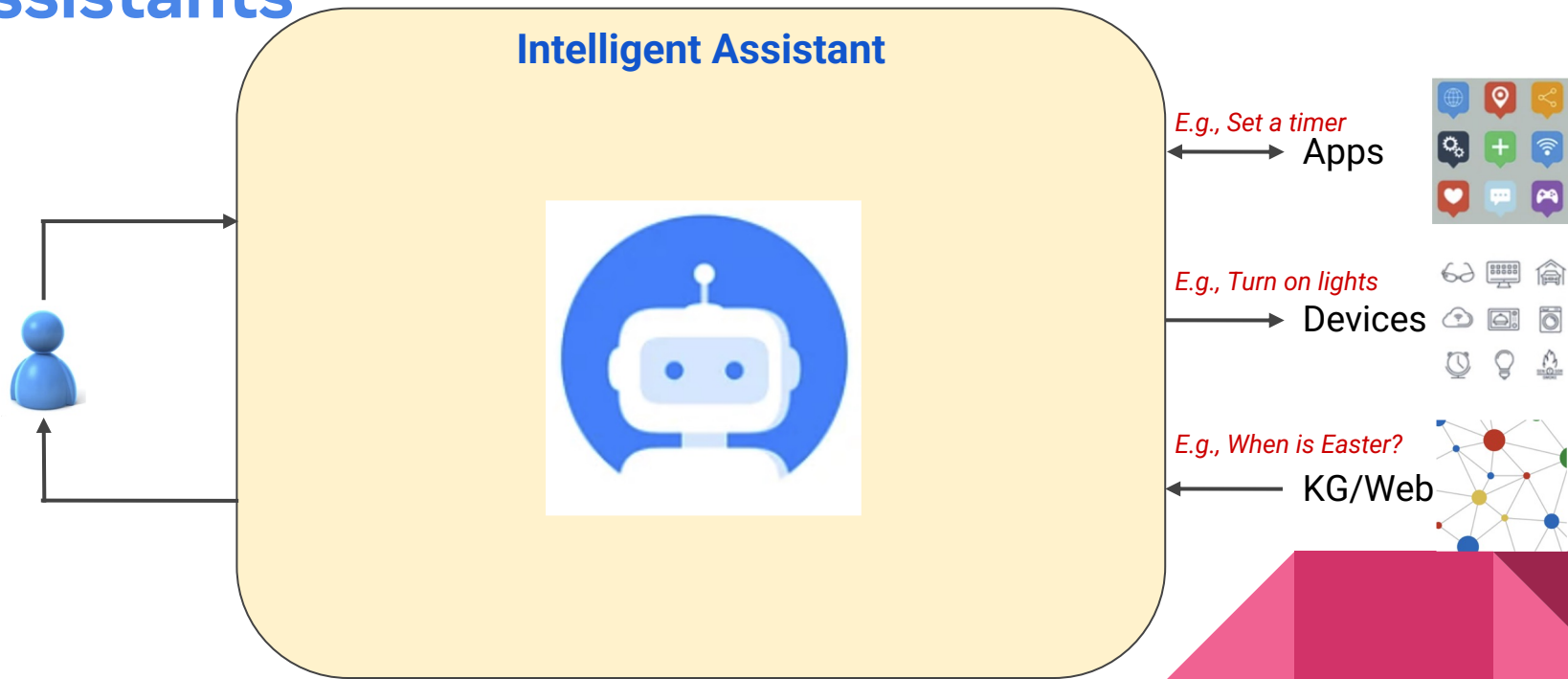
## Goals

- Introduce you the interesting research problems for Intelligent Assistants
  - Invite you to open new doors to build next-generation Intelligent Assistants
- 

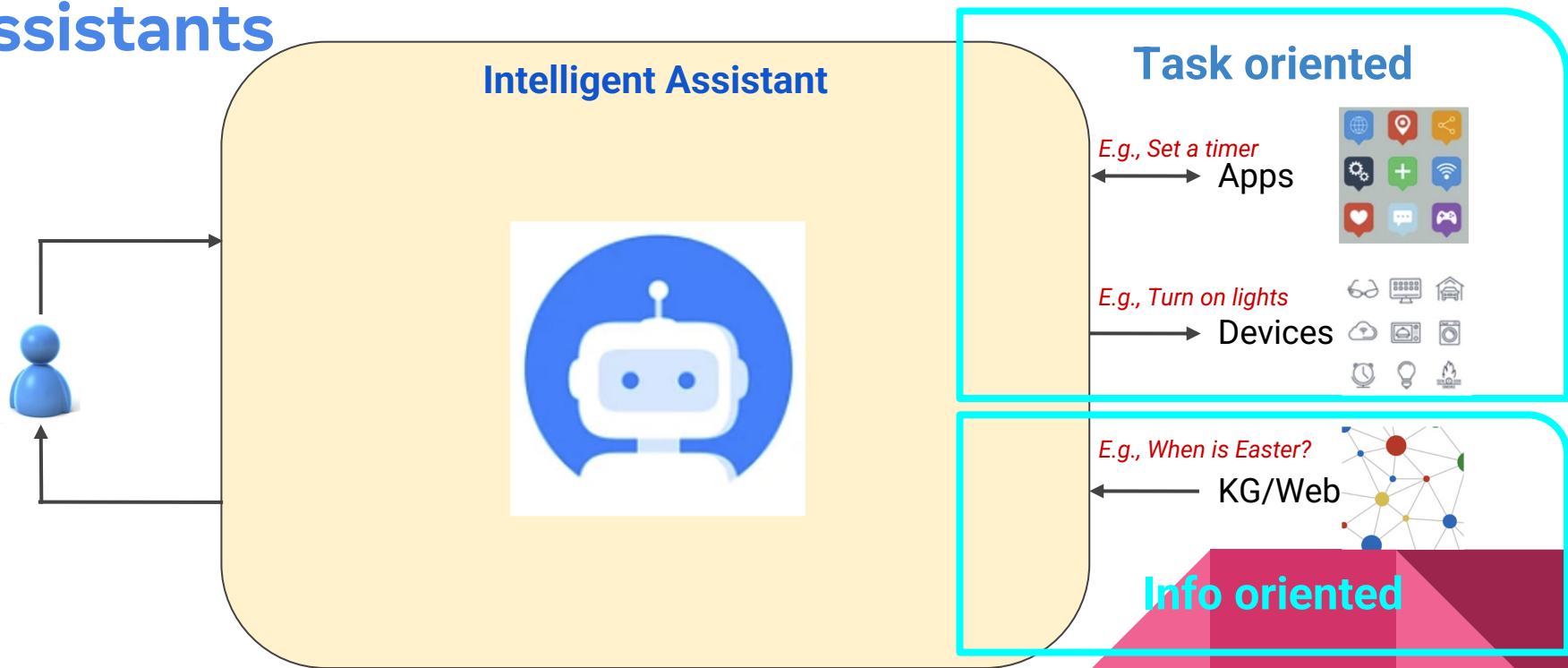


# **Technologies to Support Current Intelligent Assistants**

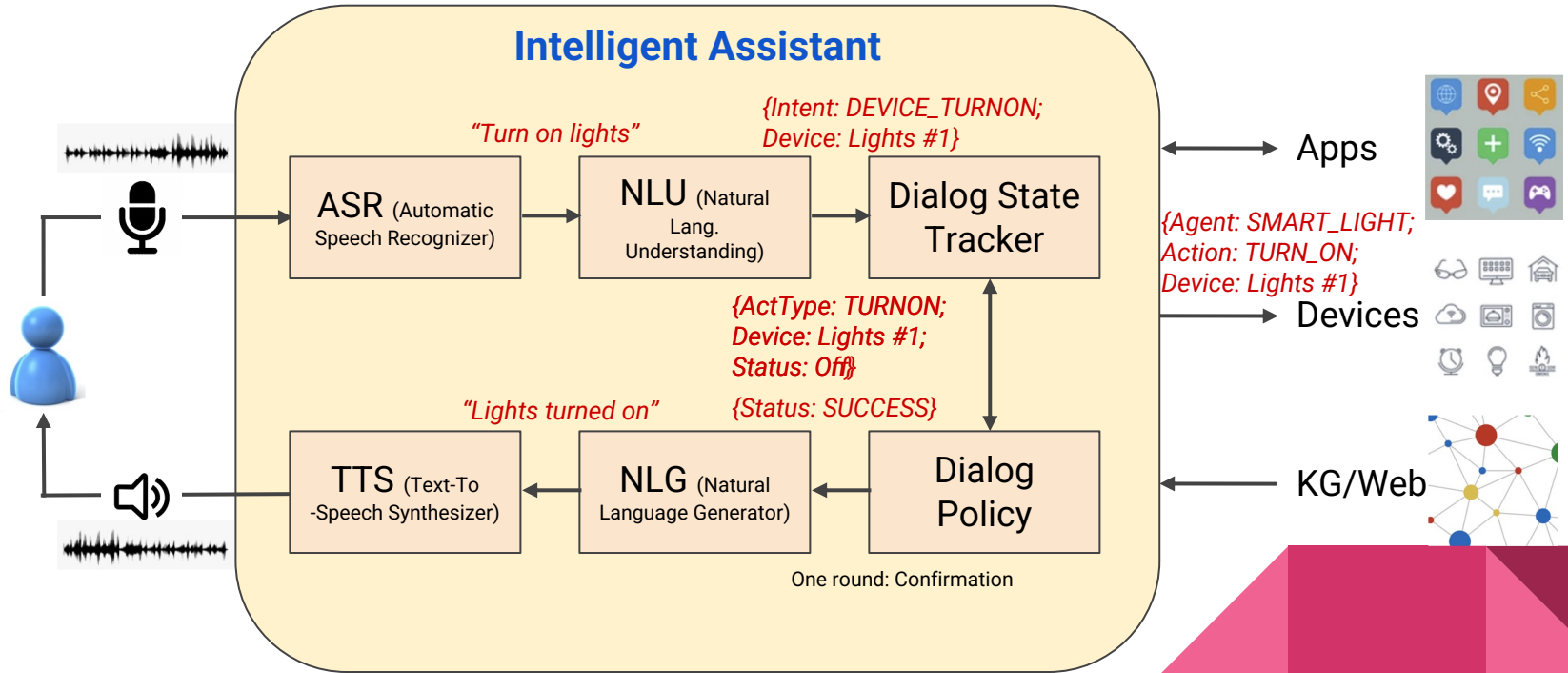
# Two Types of Commands to Intelligent Assistants



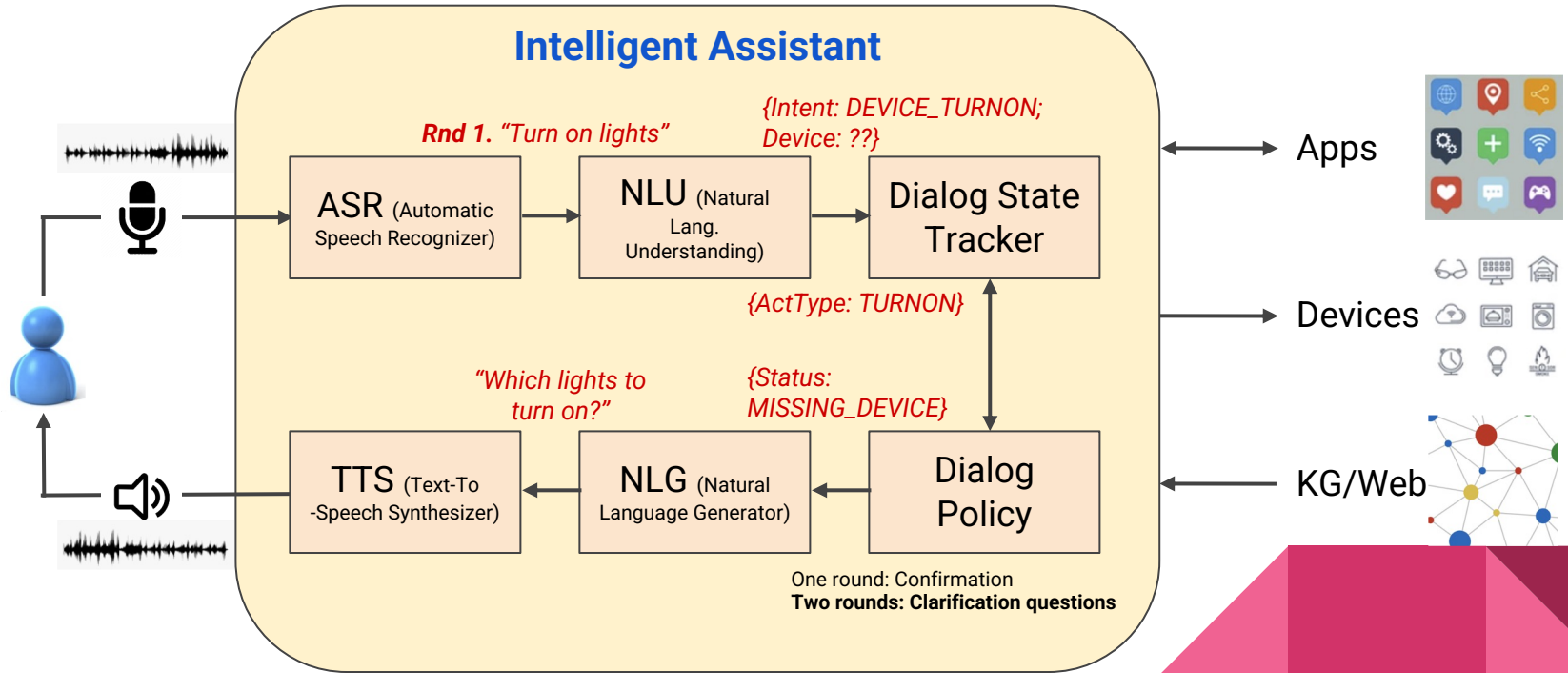
# Two Types of Commands to Intelligent Assistants



# Int. Asst. Is Essentially a Conversation System

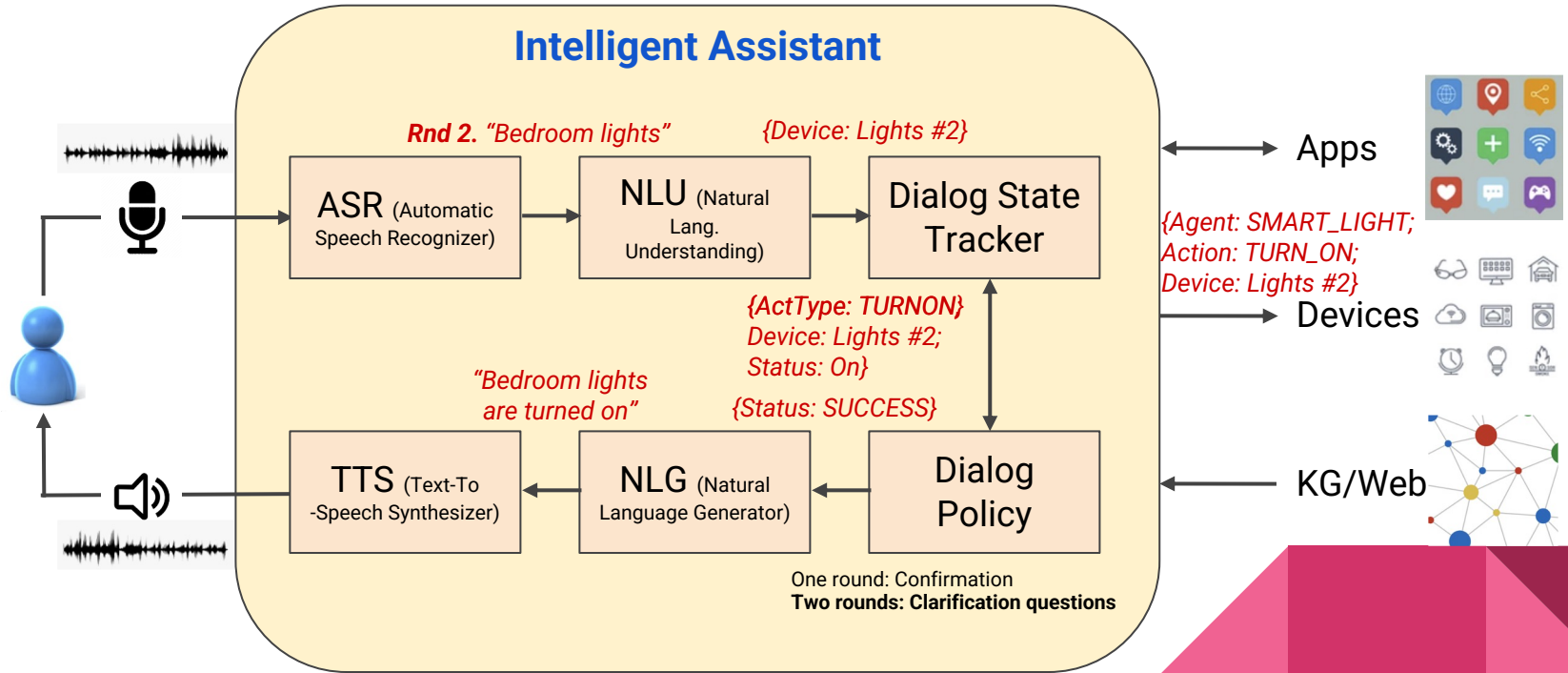


# Int. Asst. Is Essentially a Conversation System

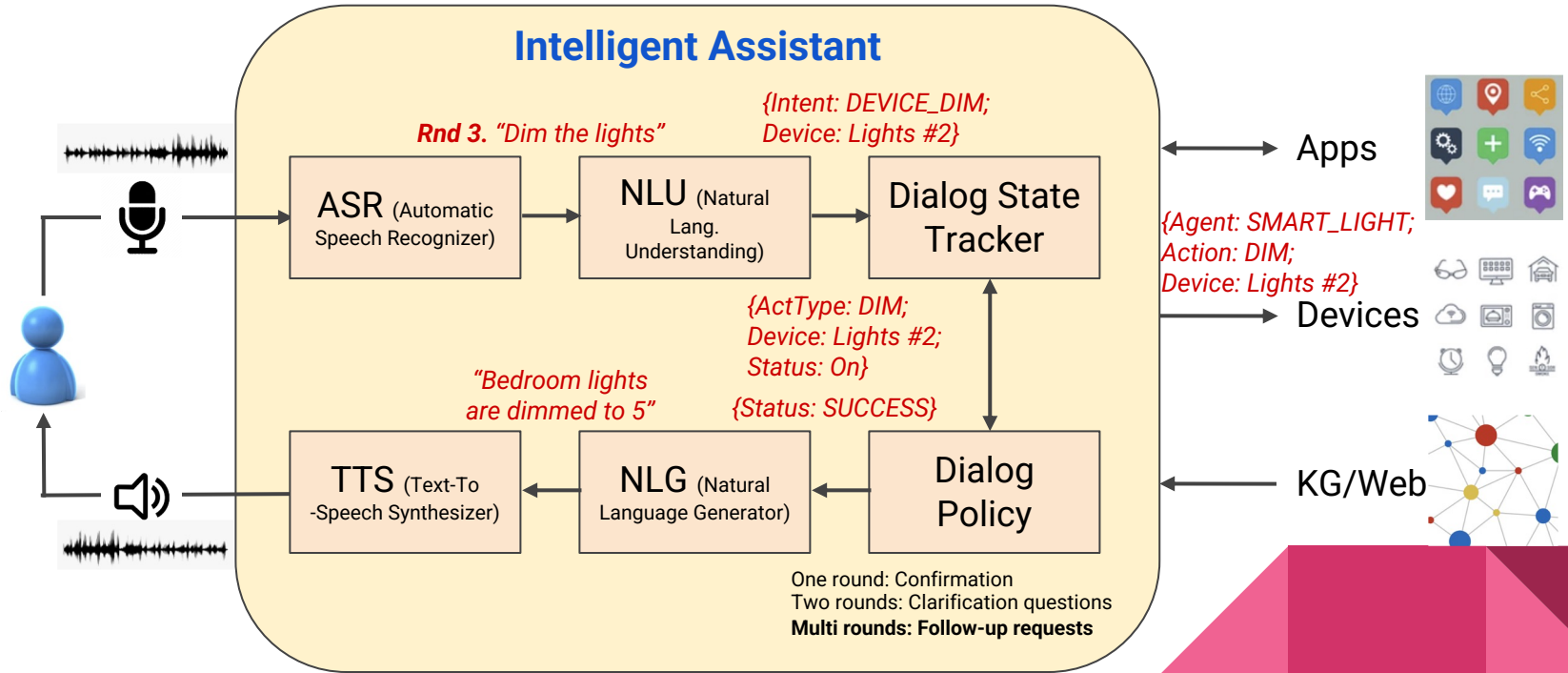




# Int. Asst. Is Essentially a Conversation System

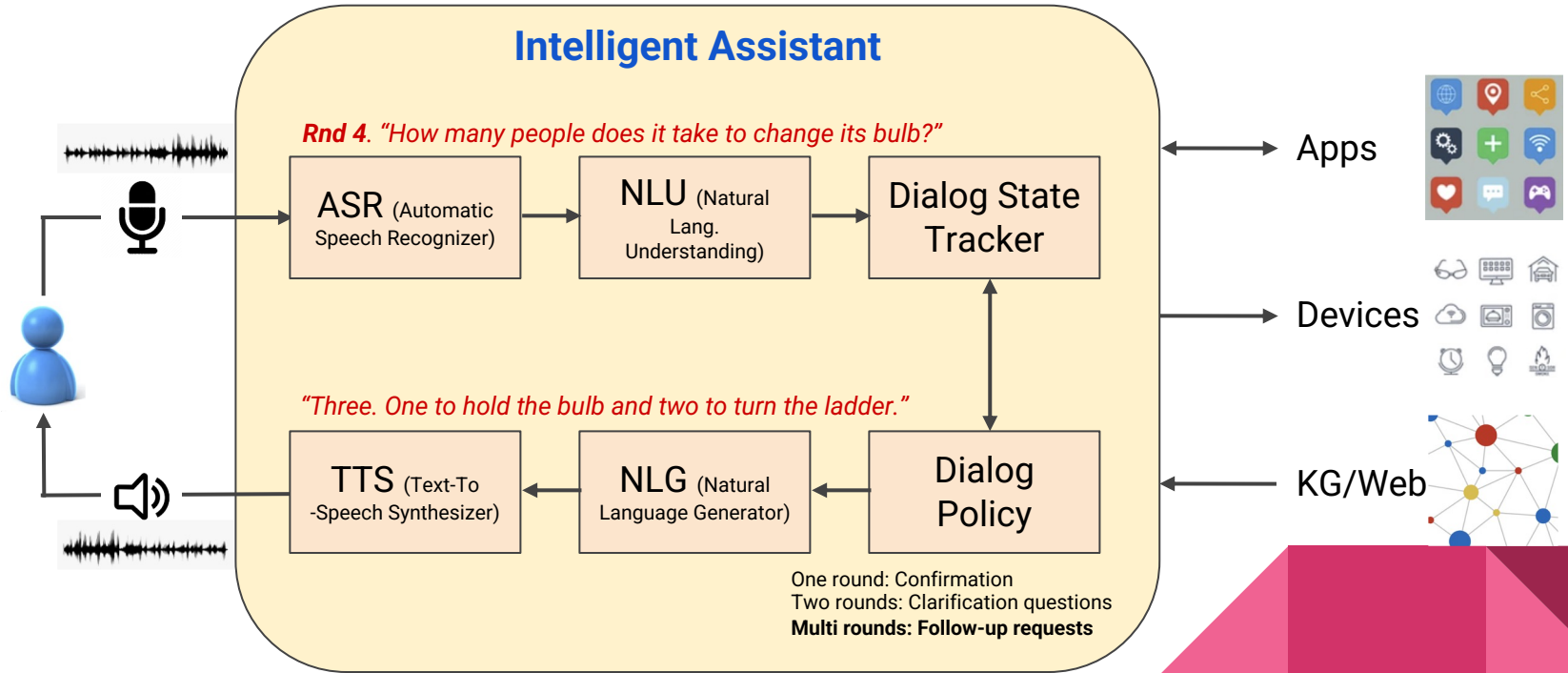


# Int. Asst. Is Essentially a Conversation System

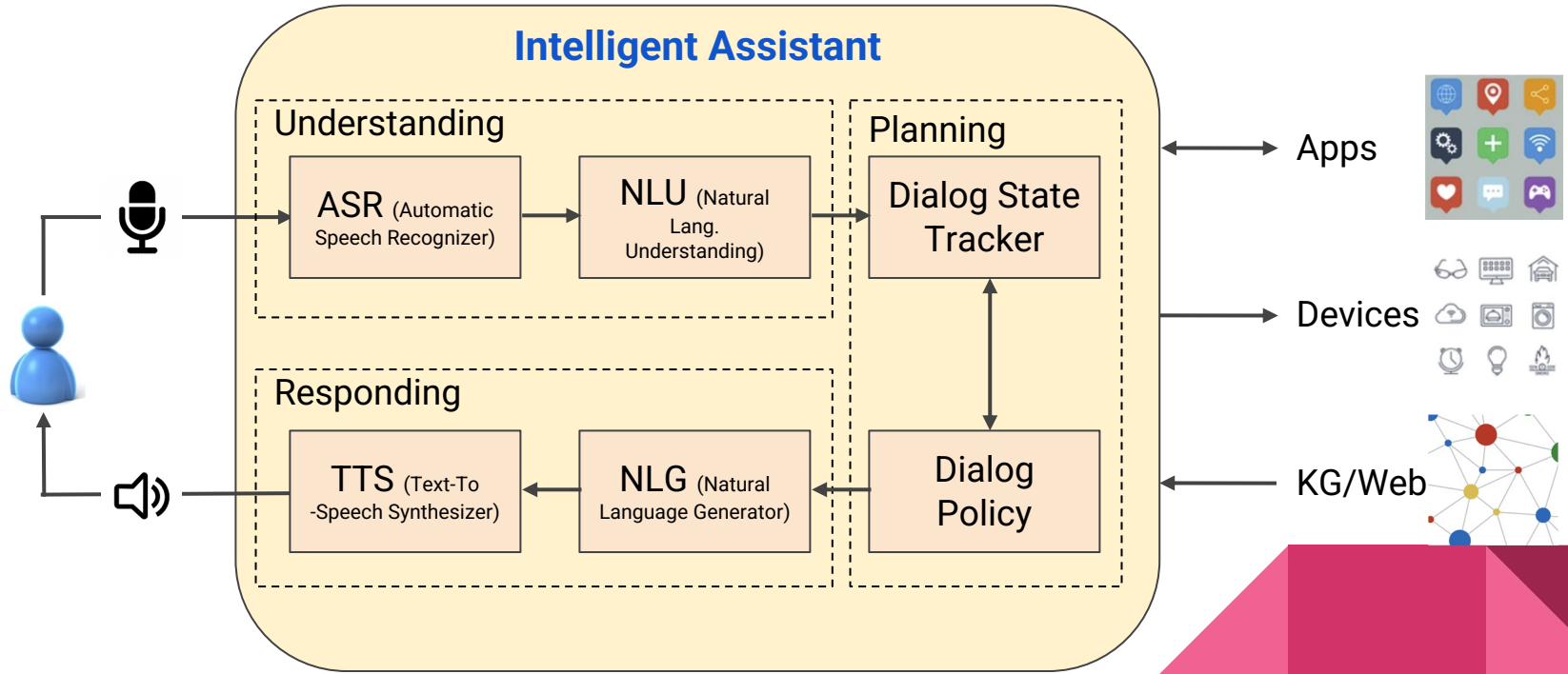


One round: Confirmation  
Two rounds: Clarification questions  
Multi rounds: Follow-up requests

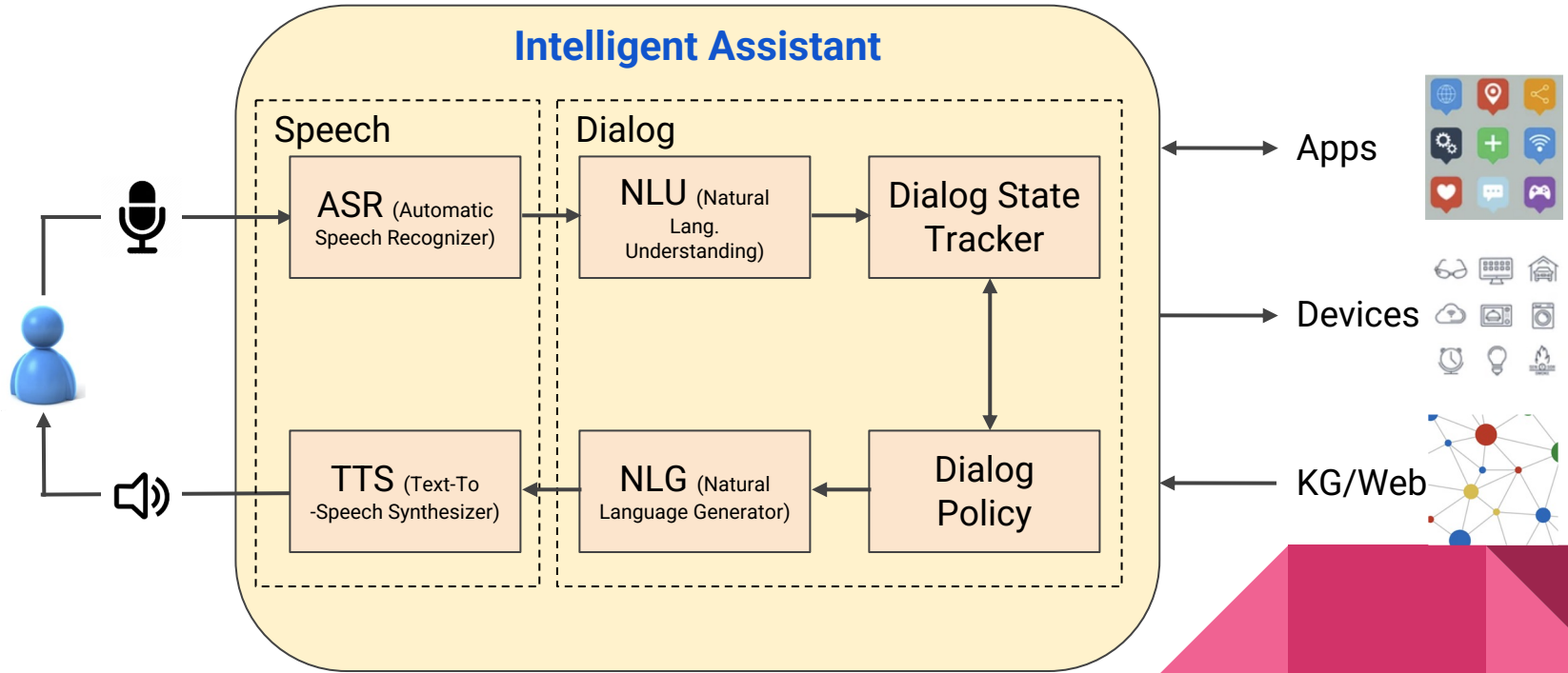
# Int. Asst. Is Essentially a Conversation System



# Int. Asst. Is Essentially a Conversation System

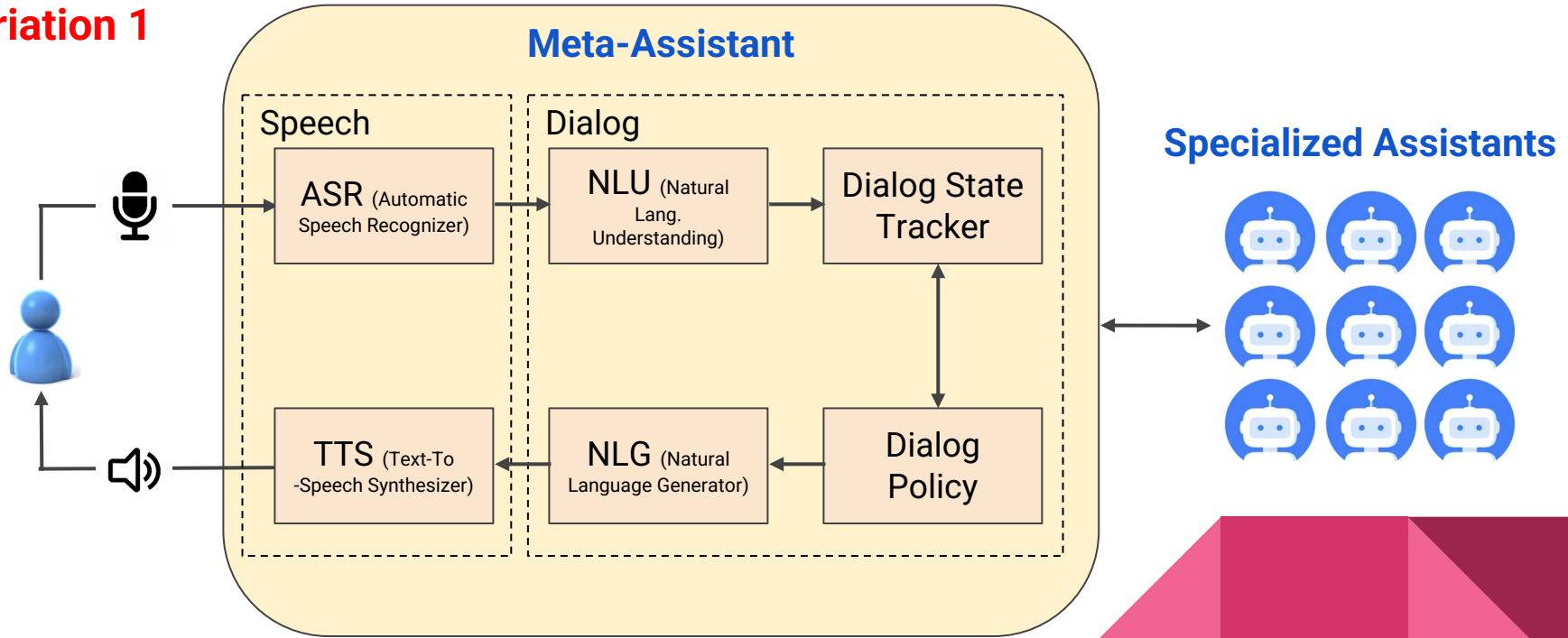


# Int. Asst. Is Essentially a Conversation System



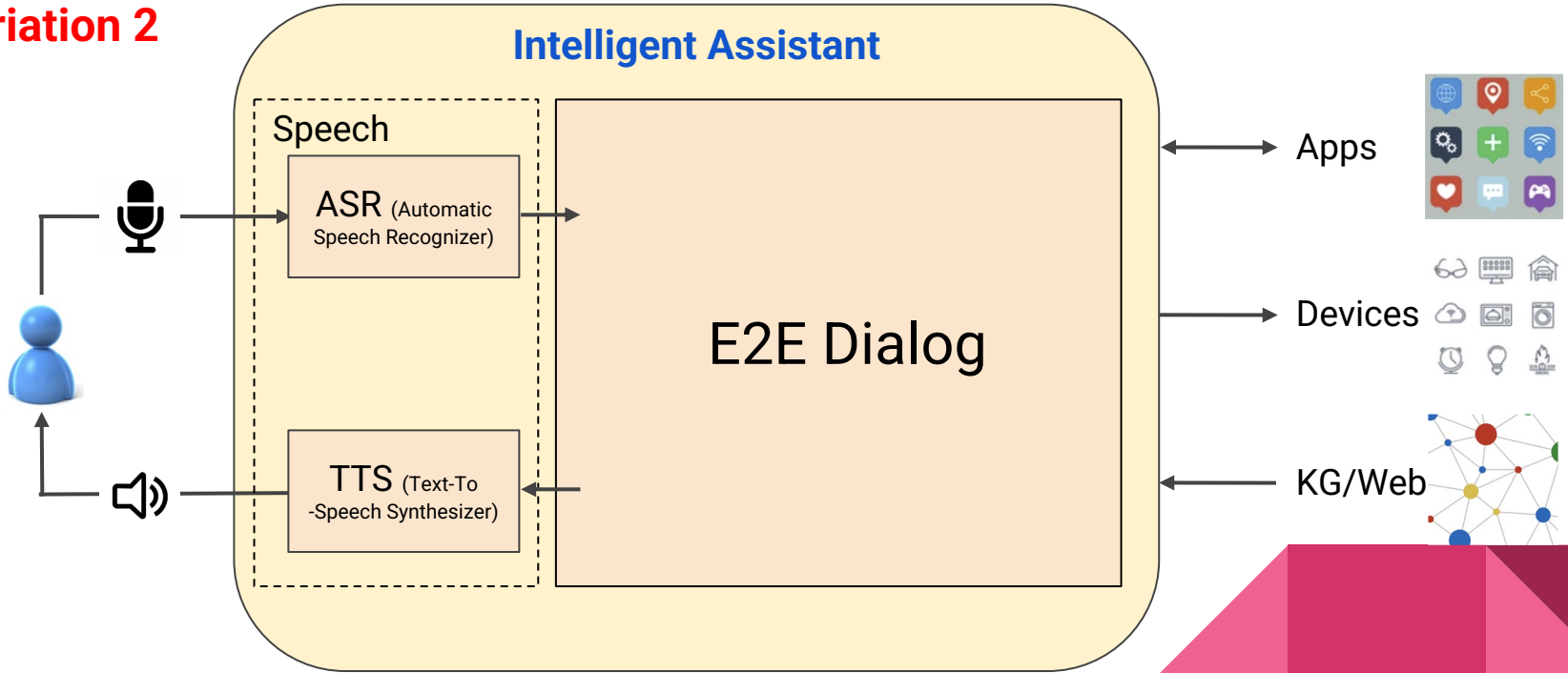
# Int. Asst. Is Essentially a Conversation System

## Variation 1

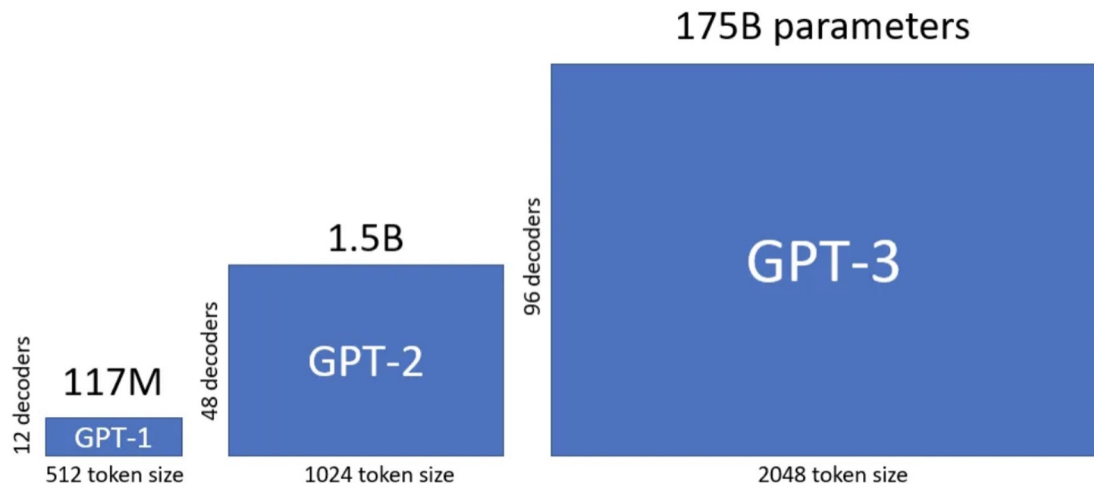


# Int. Asst. Is Essentially a Conversation System

## Variation 2



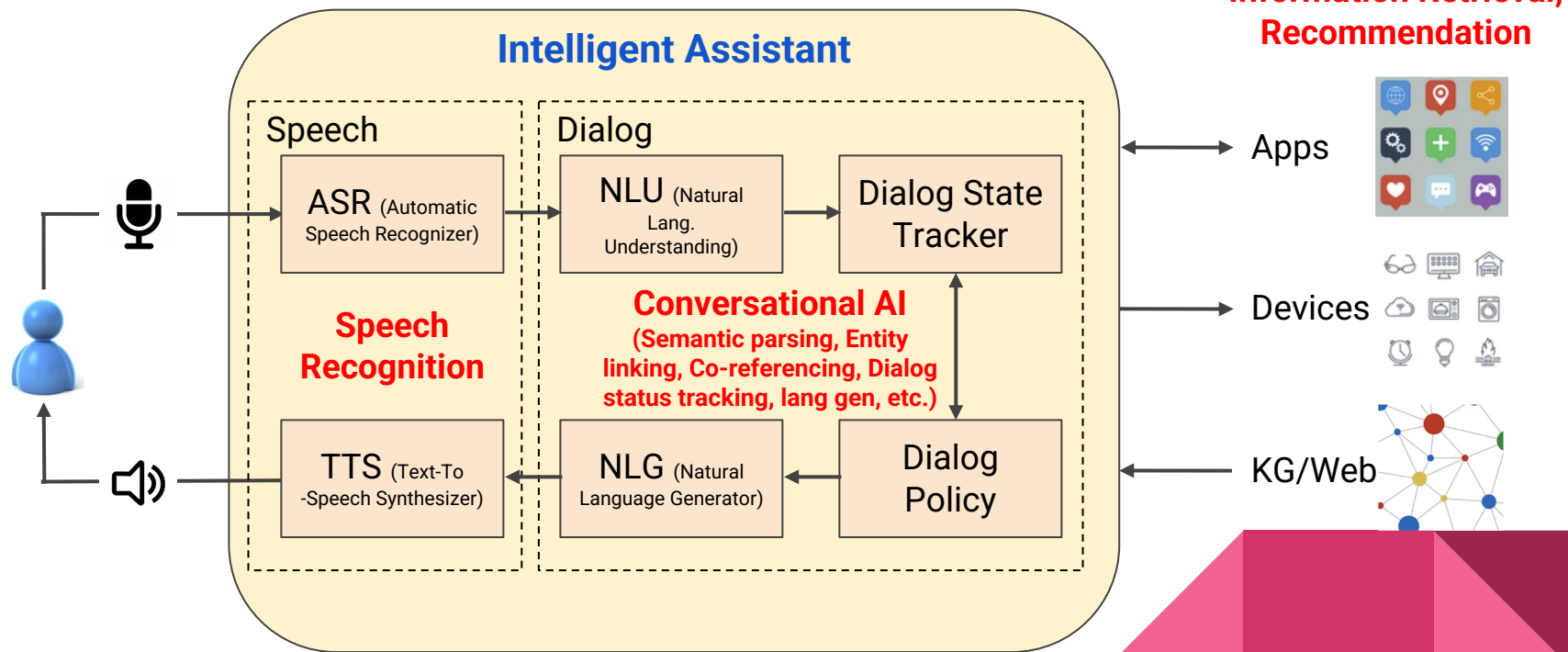
# GPT for Informational Oriented Assistant



- [GPT](#): Pre-training + in-domain fine-tuning
- [GPT 2](#): 1) Increase volume of training data;  
2) Describe tasks as a part of the input
- [GPT 3](#): Increase model size to enable few-shot learning



# Related Research Areas



# What Can Be Improved?



- How to increase accuracy?
- How to allow easy scale-up to new tasks, new domains, and new languages?
- How to make the multi-turn conversations smoother?
- How to make the assistants a know-it-all?
- How to provide better content in search, QA, recommendation?
- How to reduce response latency?
- etc.

# Ideal Assistant Revisited–Missing Pieces

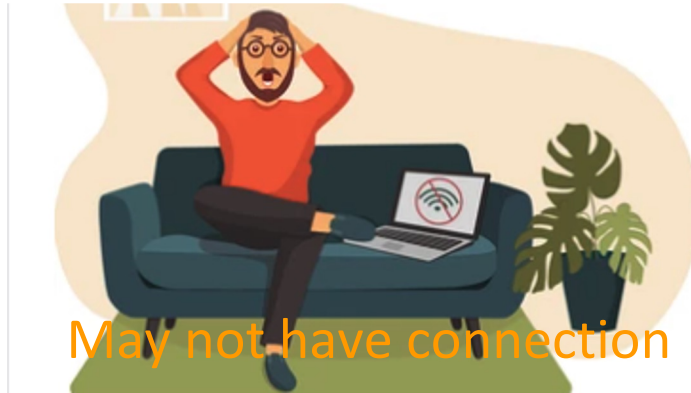
An *intelligent assistant* should be an agent that **knows you and the world**, can **receive your requests** or **predict your needs**, and provide you the **right services at the right time** with your permission.





# **Challenges and Initial Solutions to Next-Generation AR/VR Assistants**

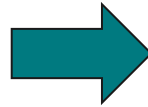
# What Is Different for An AR/VR Assistant?



# From Voice-Only to Multi-Modal



“How tall is Empire State Building?”

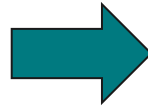


“What’s the name of this building and how tall is it?”

# From Context-Agnostic to Context-Aware



“Show my shopping list”

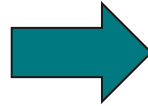


“Remember to buy apples and bananas at the grocery store around the corner”

# From Reactive to Proactive



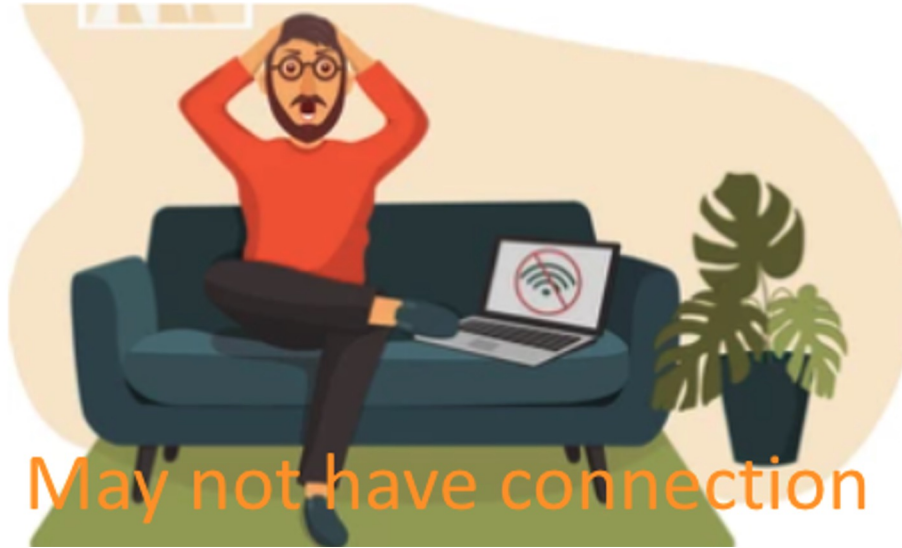
“What’s the weather today?”



“Today is sunny, 70 degree. Would you like to play your favorite morning music?”



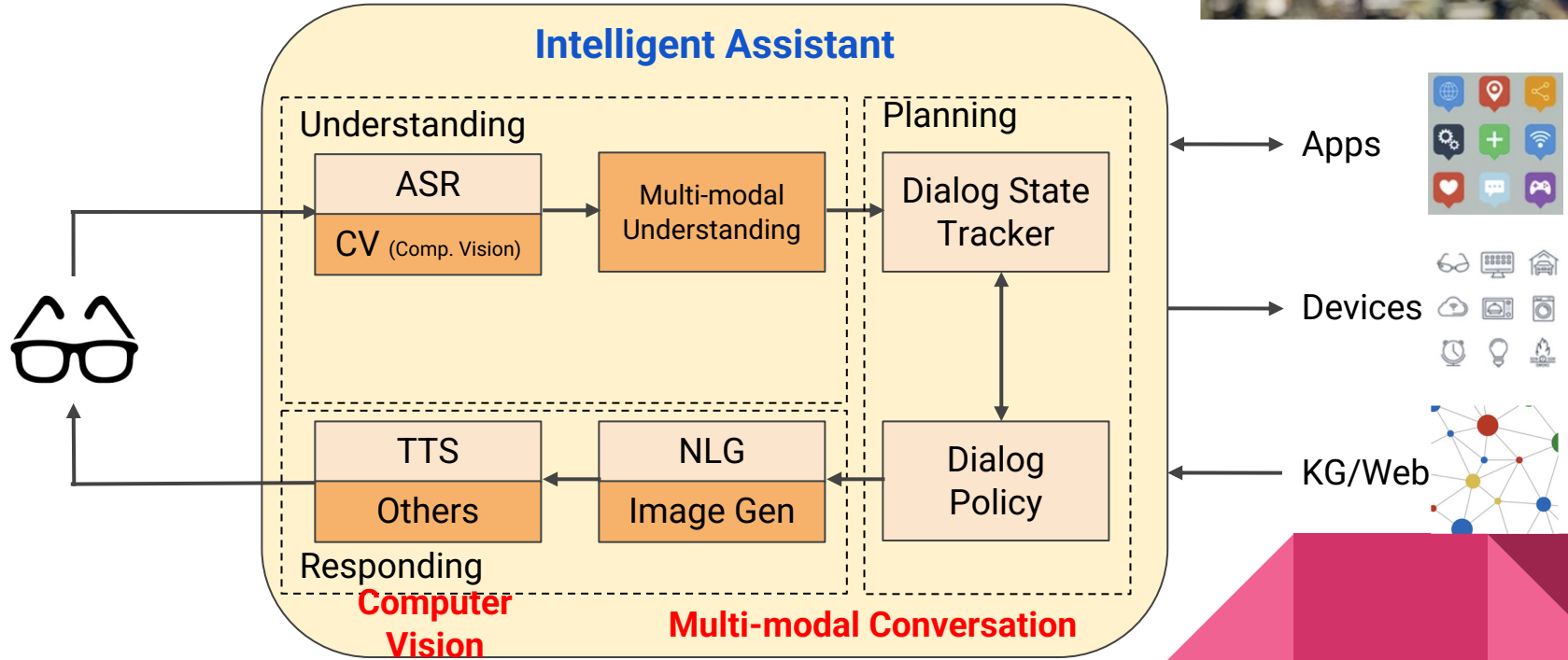
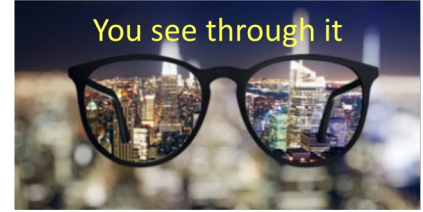
# From Server-Side to On-Device



May not have connection

+ Privacy!!!

# Direction 1. Multi-Modal Assistant



# 1-1. Multi-Modal—EgoOCR



Normal (By cell phone)



## Common challenges

- Arbitrary shape
- Complex layout
- Multiple languages

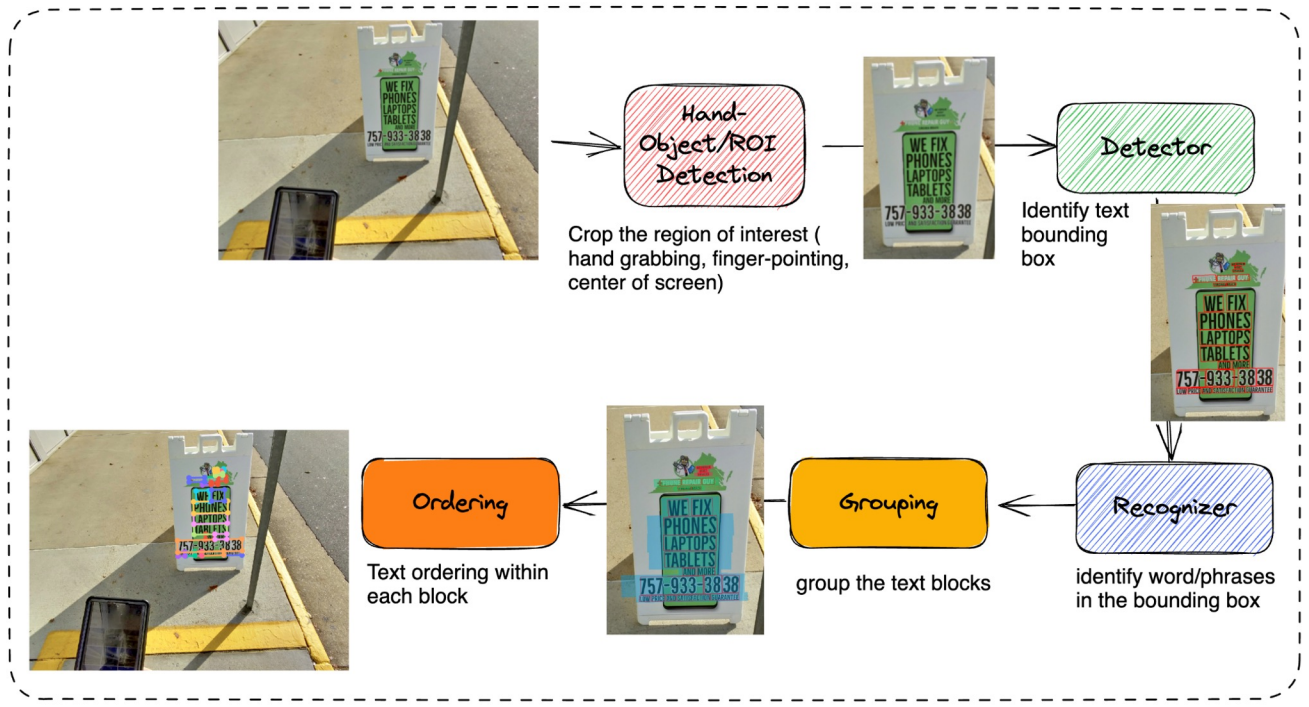
Ego-centric (By glasses)



## Unique challenges

- Text too small or too large
- Sub-optimal lighting cond
- Blurry images by movement
- User focus unclear
- Non-horizontal direction
- On-device

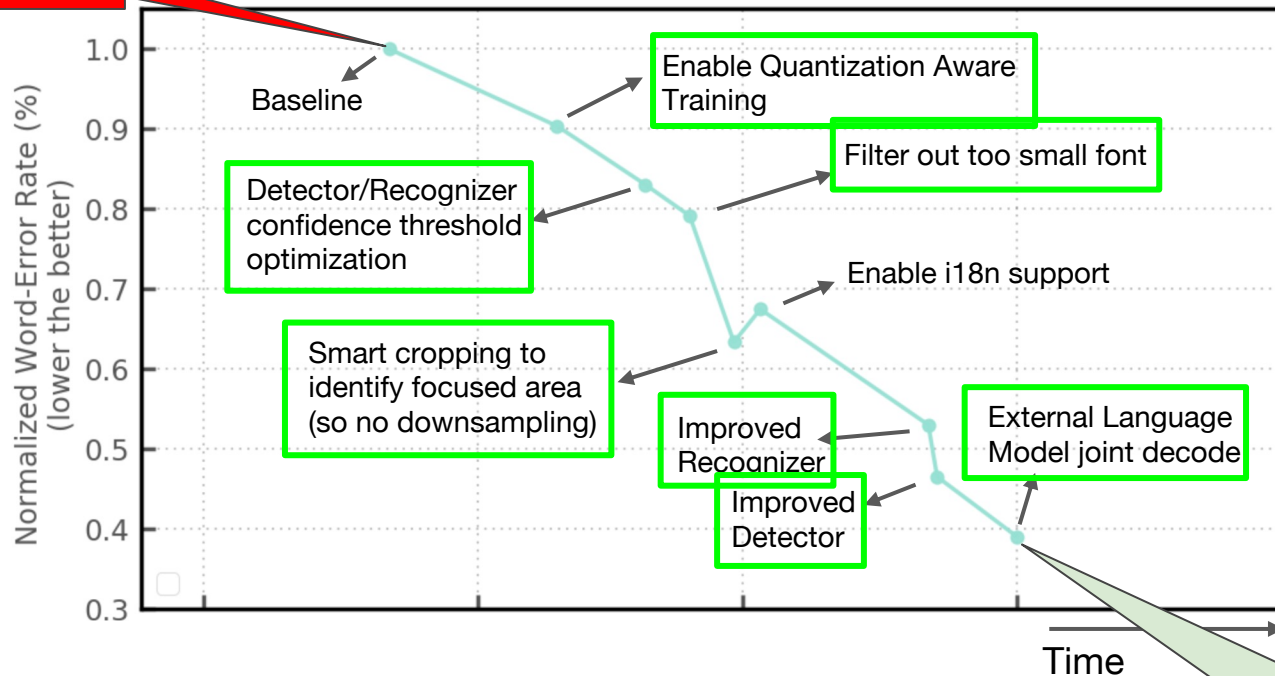
# 1-1. Multi-Modal—EgoOCR



# 1-1. Multi-Modal—EgoOCR

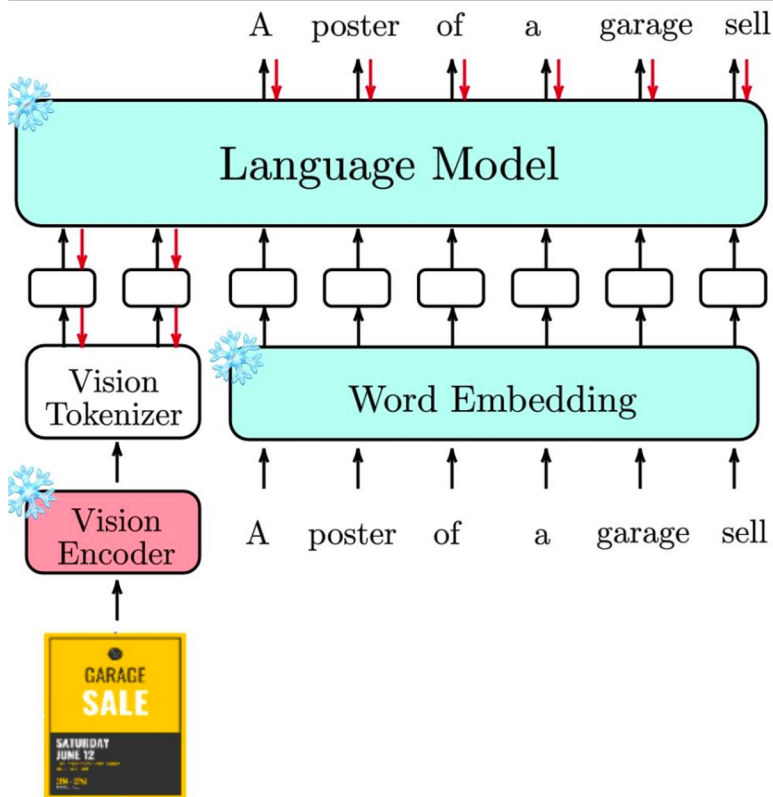


60% Absolute WER



25% Absolute WER

## 1.2. Multi-Modal—LLM





### Key advantages

- On top of any state-of-the-art LM; e.g., LLaMA, ChatGPT.
- Allowing for *combinatory* modalities as conditioning context
- Working for *resource-scarce* modalities; e.g., audio



## 1.2. Multi-Modal—LLM



Image Input	Text Prompt Input	MM-LLM Generated Output
	<p><i>Should I attend this event?</i></p> <p><i>Let's think step by step.</i></p>	<p>The event is a garage sale. Garage sales are a great way to save money. So the final answer is yes.</p>
	<p><i>Is this soup healthy (looking at ingredients)?</i></p> <p><i>Let's think step by step.</i></p>	<p>The soup is made with carrots, celery, onions, and carrots. So the answer is yes.</p>

## 1.2. Multi-Modal—LLM





Image Input	Text Prompt Input	MM-LLM Generated Output
	<p><i>Describe the scene in front of me.</i></p>	<p>A man is standing in front of a bunch of fruits and vegetables.</p>
	<p><i>If I were a renowned art critic, how would I describe this painting?</i></p>	<p>It is a large painting with a lot of detail.</p>





## 1.2. Multi-Modal—LLM



Audio Input	Text Prompt Input	MM-LLM Generated Output
 <a href="#">&lt;frying.wav&gt;</a>	<p><b>Where am I?</b></p> <p><i>Let's think step by step. First, describe the sound:</i></p>	<p>The sound of cooking food. A kitchen is a place where food is cooked. So, the answer is a kitchen.</p>
 <a href="#">&lt;seagull.wav&gt;</a>	<p><b>What animal is this?</b></p> <p><i>Let's think step by step.</i></p>	<p>A crow is a bird that makes a high-pitched sound.</p>



## 1.2. Multi-Modal—LLM

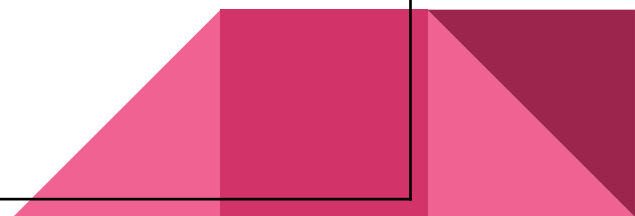


Audio Input	Text Prompt Input	MM-LLM Generated Output
 <a href="#">&lt;rain_thunder.wav&gt;</a>	<p><b><i>What's the weather like now?</i></b></p> <p><i>Let's think step by step. First, describe the sound:</i></p>	<p>The sound of a rainstorm. The weather now is rainy. So, the answer is rainy.</p>
 <a href="#">&lt;birds.wav&gt;</a>	<p><b><i>Describe the sound:</i></b></p>	<p>Birds chirping</p>

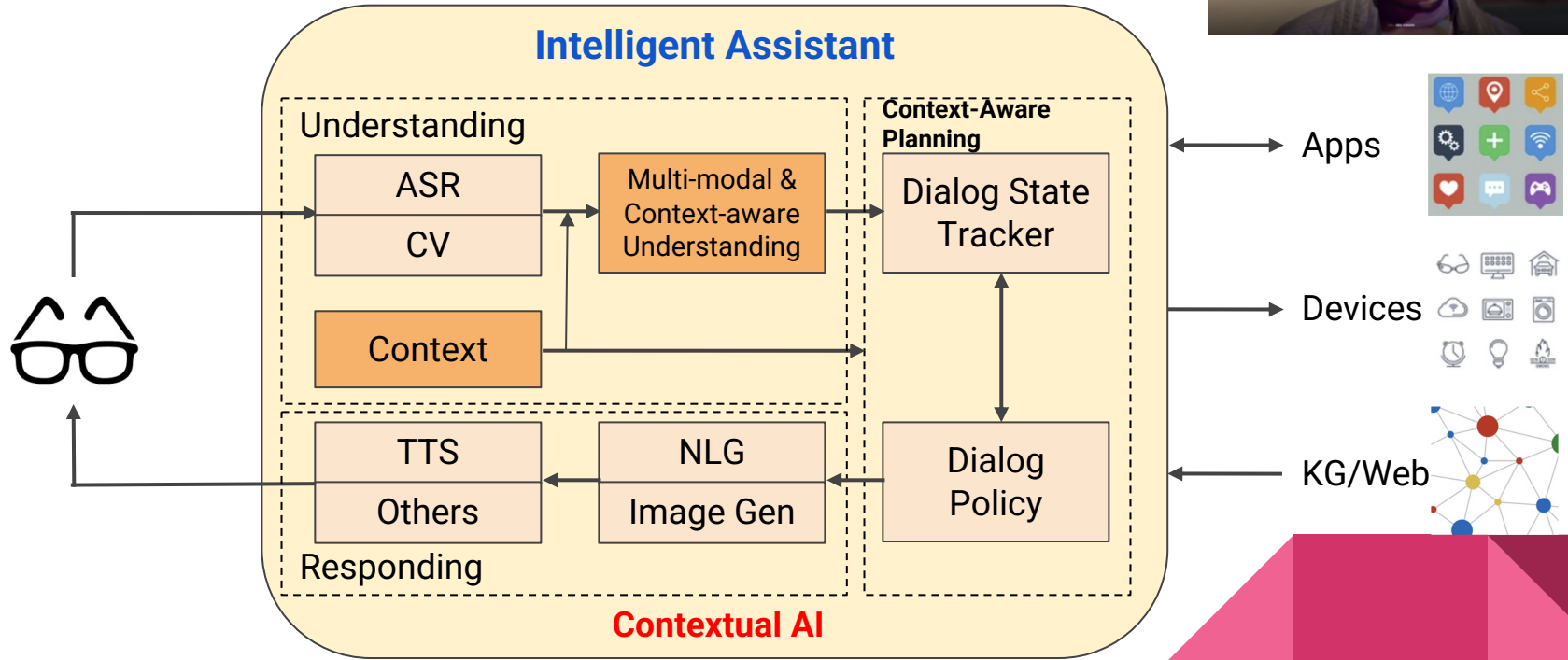
## 1.2. Multi-Modal—LLM



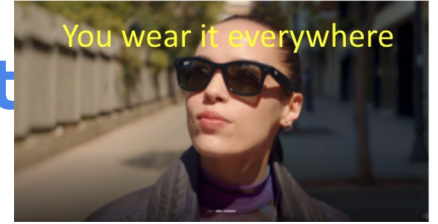
Video Input	MM-LLM Generated Output
	#o man x picks cards from the table
	#c c rinses the bowl with water from the tap



# Direction 2. Context-Aware Assistant



## Direction 2. Context-Aware Assistant



- Context-aware assistants
  - examine your surroundings, and
  - use this context to personalize a product experience.



# Direction 2. Context-Aware Assistant



- Context-aware assistants
  - examine your surroundings, and
  - use this context to personalize services

## Context

(Time, Location, Scene,  
Activity, Event, etc.)

When is it?

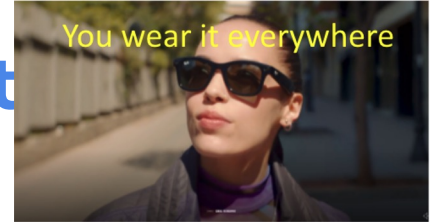
Where are you?

What are you doing?

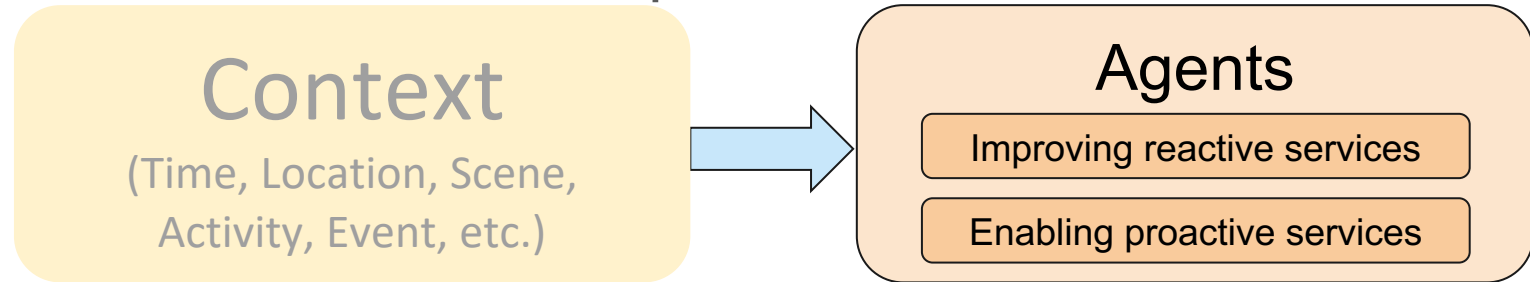
Whom are you together with?

What's surrounding you?

# Direction 2. Context-Aware Assistant



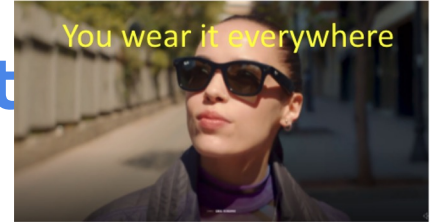
- Context-aware assistants
  - examine your surroundings, and
  - use this context to personalize services



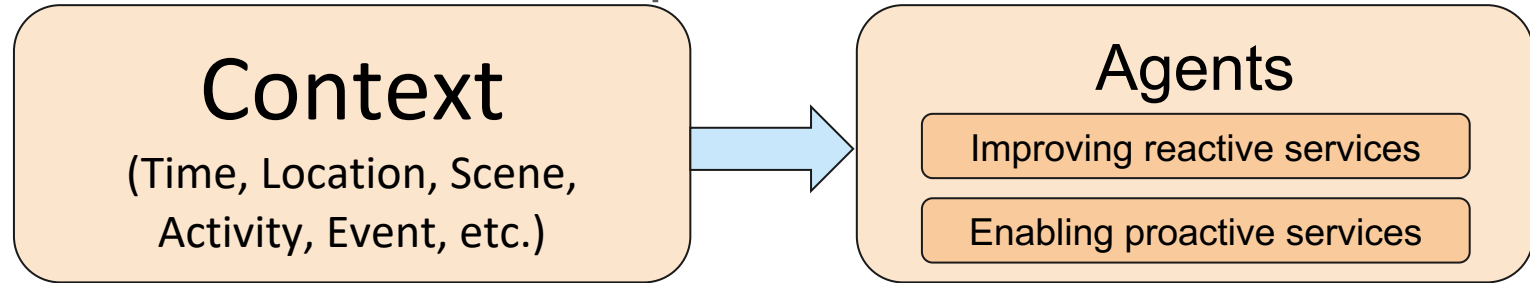
When is it?  
Where are you?  
What are you doing?  
Whom are you together with?  
What's surrounding you?

Context-aware ranking  
Contextual recommendation  
Contextual reminder, etc.

# Direction 2. Context-Aware Assistant



- Context-aware assistants
  - examine your surroundings, and
  - use this context to personalize services



When is it?

Where are you?

What are you doing?

Whom are you together with?

What's surrounding you?

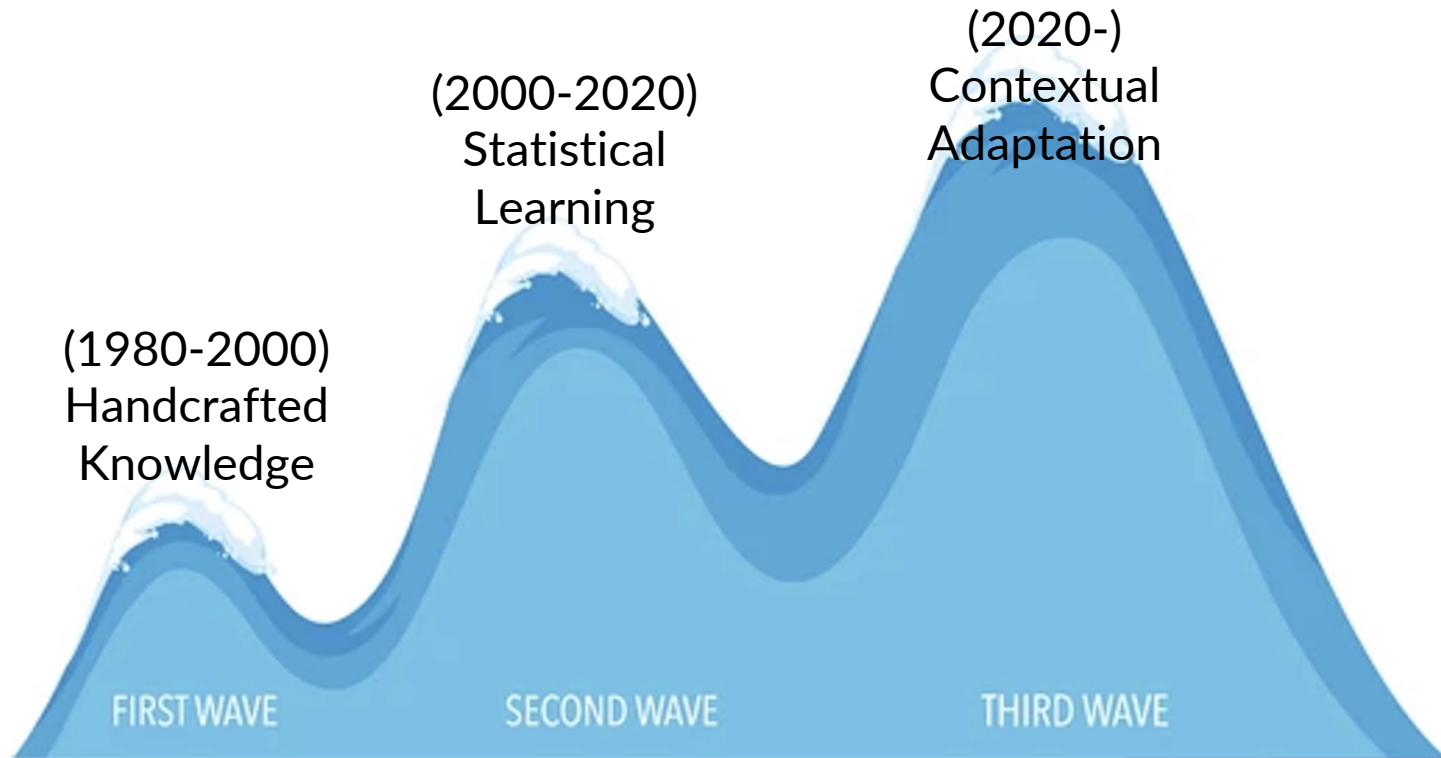
Context-aware ranking

Contextual recommendation

Contextual reminder, etc.



# Direction 2. Context-Aware Assistant



Reference: [A DARPA Perspective on Artificial Intelligence](#)

## 2.1. Business Recognition

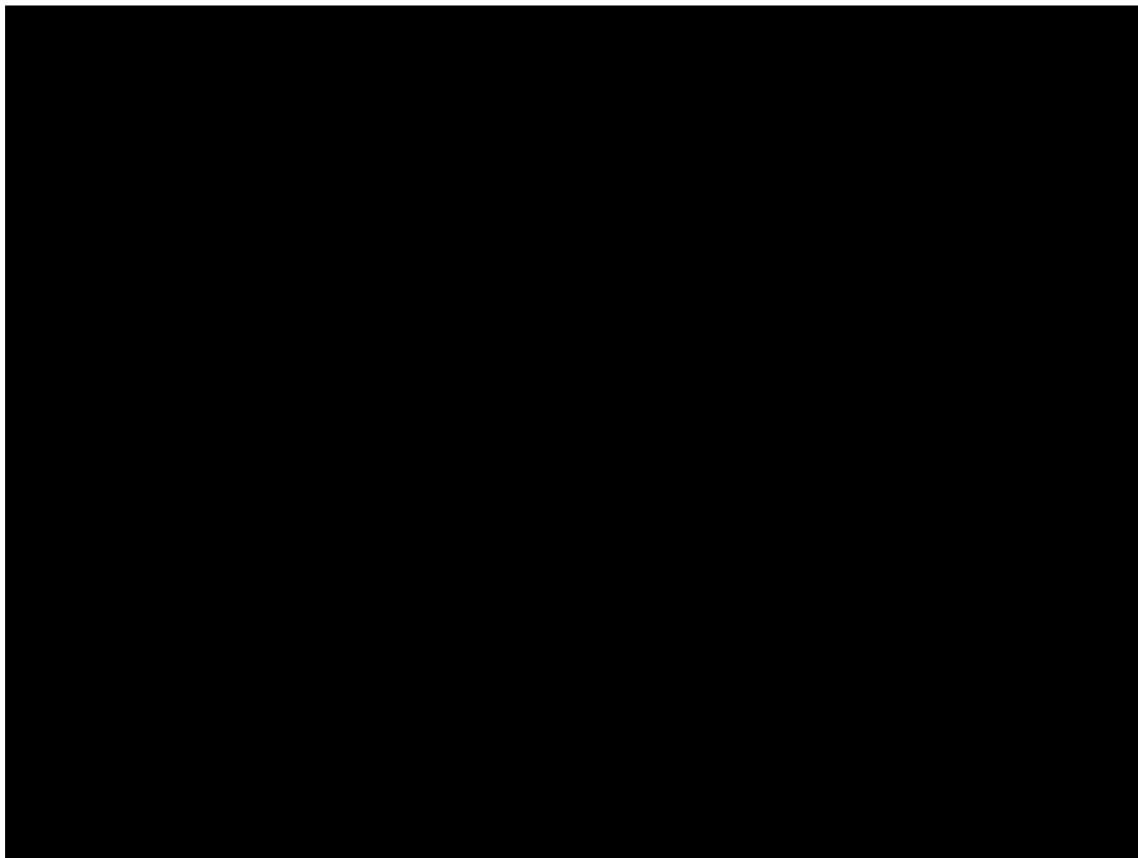


Method	hit@1
Text	67%
Text+Image	74%
Text+Image (GPS, 30m)	86%
Text+Image (GPS, 8m)	98%



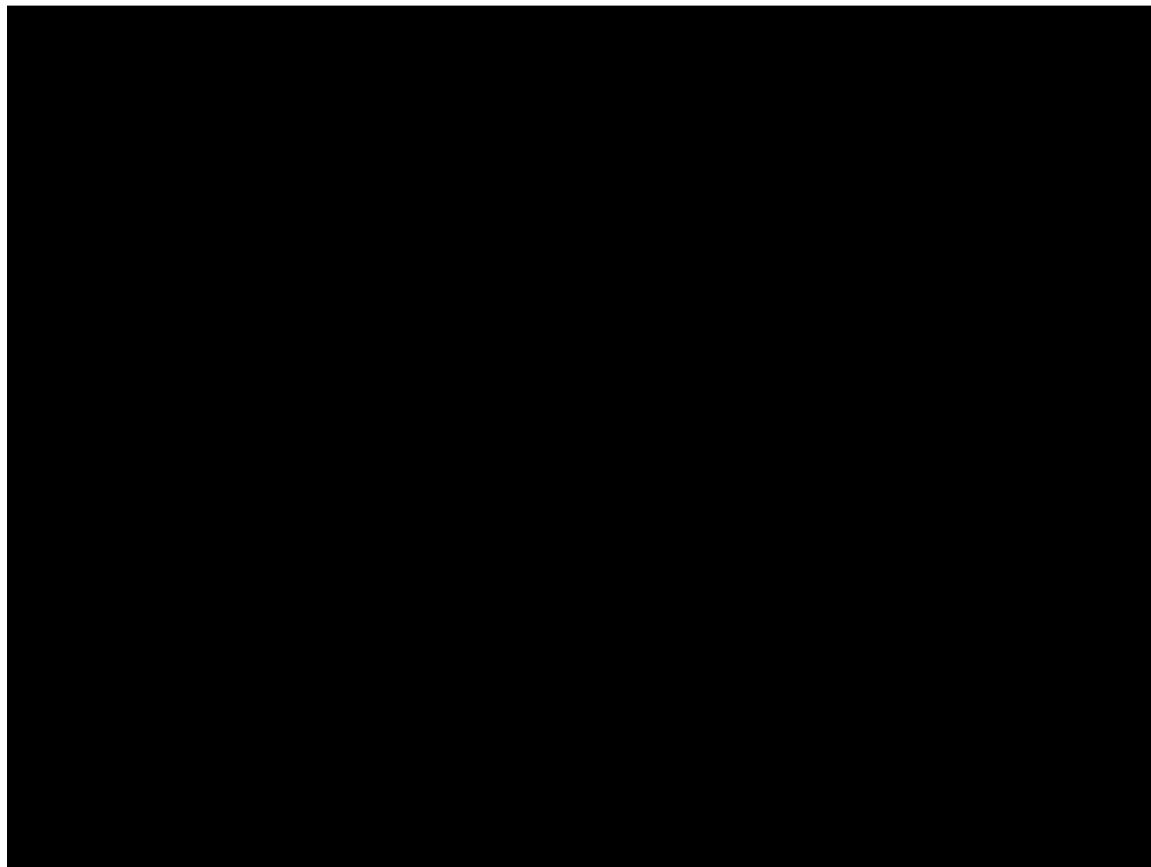
## 2.2. IMU-Based Action Detection Example 1

IMU (Inertial  
Measurement  
Unit)

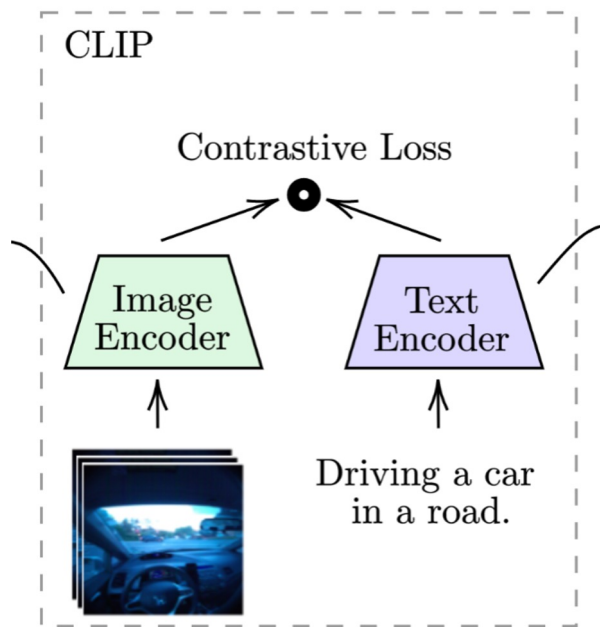


## 2.2. IMU-Based Action Detection Example 2

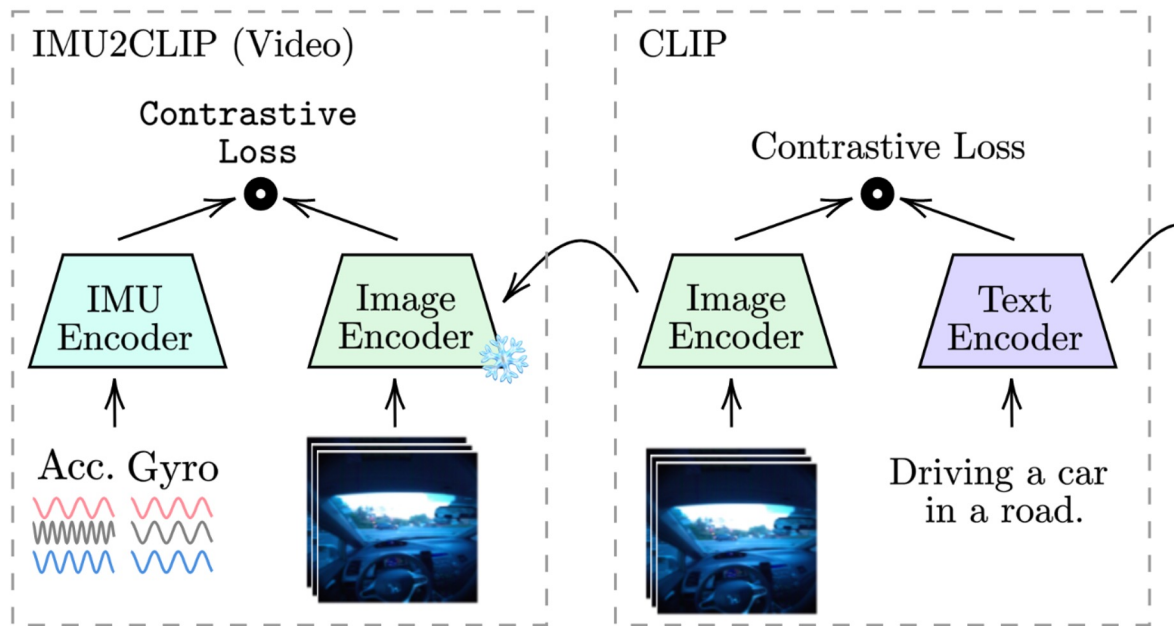
IMU (Inertial  
Measurement  
Unit)



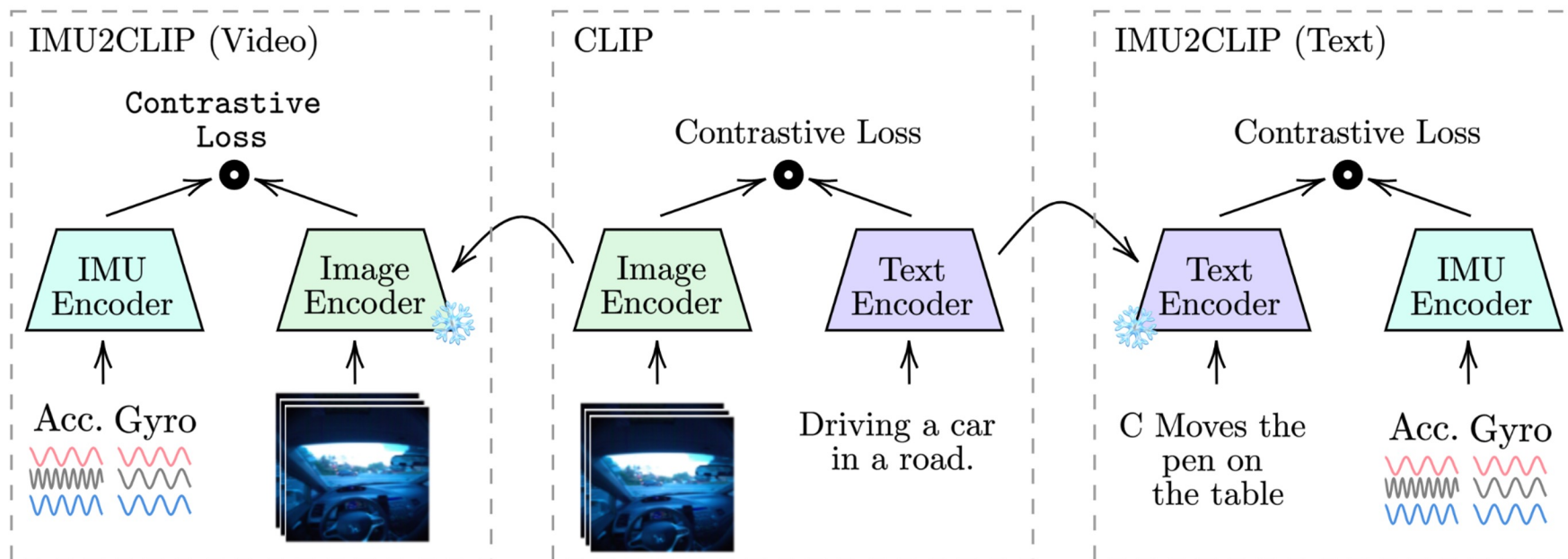
## 2.2. IMU2CLIP Aligning IMU to Video / Text



## 2.2. IMU2CLIP Aligning IMU to Video / Text



## 2.2. IMU2CLIP Aligning IMU to Video / Text



## 2.2. Aligning IMU to Video / Text

IMU2CLIP (Proposed) Modalities Used for Pre-Training			Retrieval Evaluation Metrics (Ego4D)							
			Text > IMU				IMU > Text			
IMU	Video	Text	R@1	R@10	R@50	MRR	R@1	R@10	R@50	MRR
○	○		4.17	15.62	43.06	0.0839	4.86	18.75	48.26	0.1036
○		○	7.29	28.82	60.07	0.1425	5.21	25	60.42	0.1225

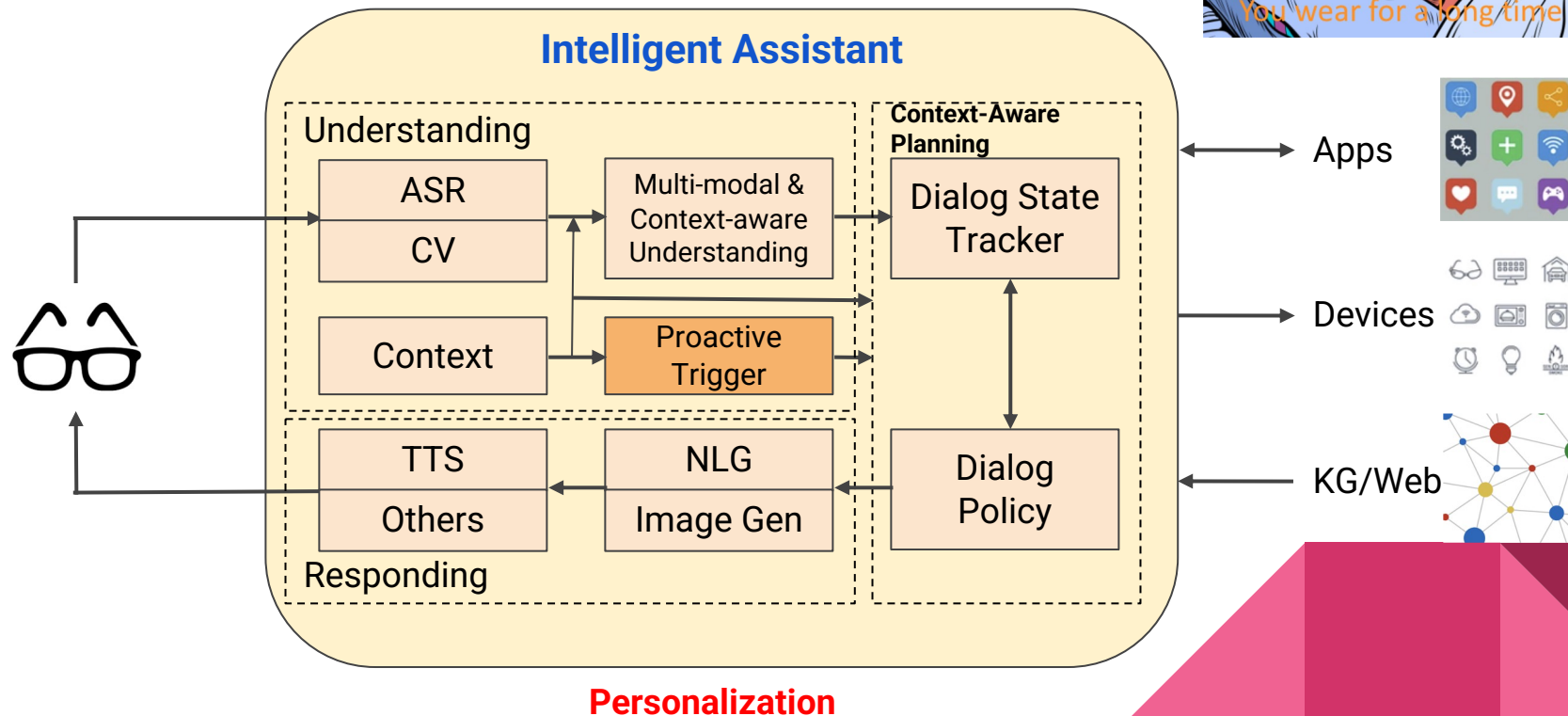
*Results:* We can detect moments & search for media based on IMU signals (using text queries)

CLIP (OpenAI)			Text > Video				Video > Text			
IMU	Video	Text	R@1	R@10	R@50	MRR	R@1	R@10	R@50	MRR
	○	○	8.33	33.68	65.28	0.1682	6.94	32.29	64.24	0.1502

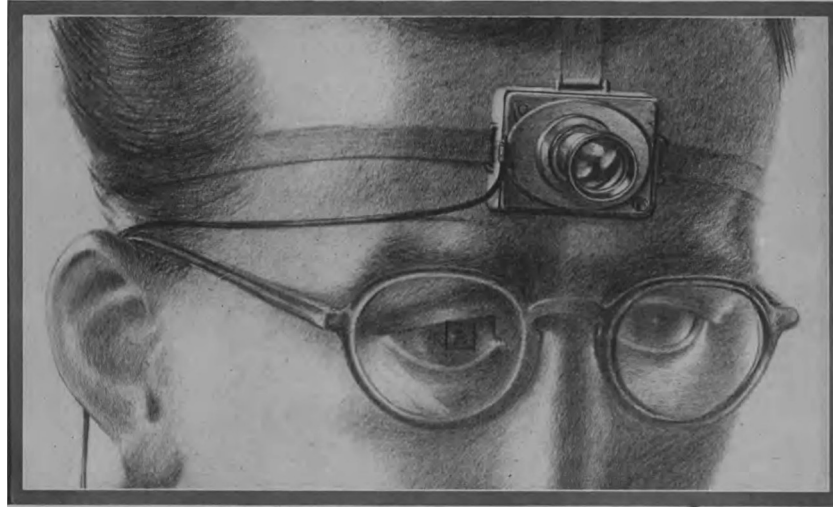
Comparable results between text <-> IMU search and Text <-> Video search



# Direction 3. Assistant Recom.



# Two Sides of One Coin (1): Great Vehicle for Life Recording



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

# Two Sides of One Coin (2): Great Vehicle for Personalized Assistance

## Utility



Where did I put my key?

I must have seen this lady before but when and where?

## Memoir



## Inspiration



At Lyon you can visit the statue of Saint-Exupéry with the Little Prince by his side. You read that book in 2018 and loved it.

# Direction 3. Conversational Recommendation

## (a) Conversational Recommendations

1 Hello, I'm looking for a good place to eat.

2 Is Downtown Phoenix a good place to start your search as you've been there a few times?

3 Yes, I'd like something there please.

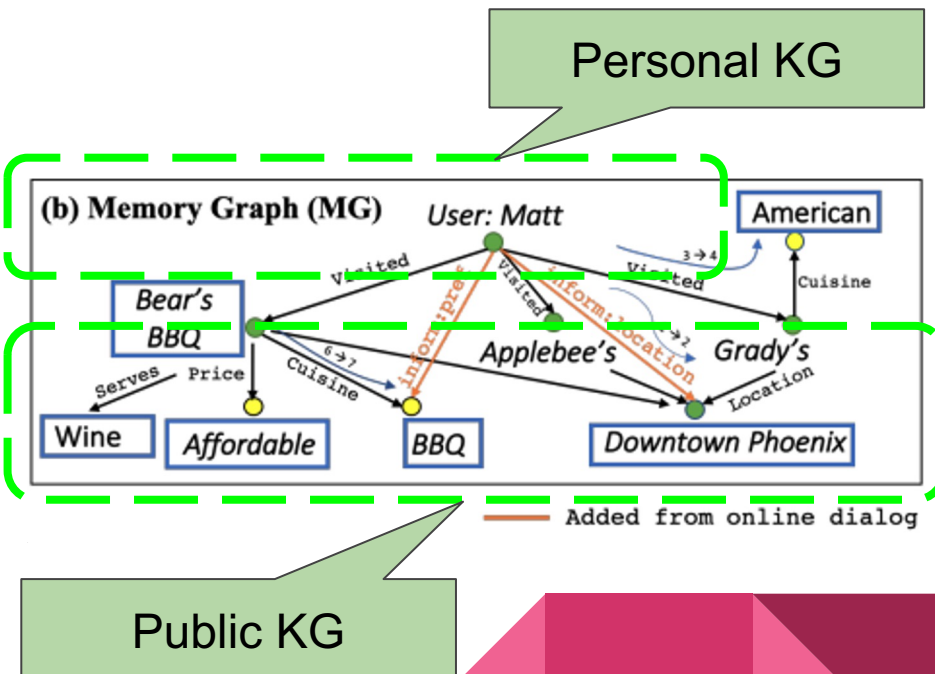
4 You ordered *American* food a few times, do you want something similar or feeling adventurous today?

6 No, I'm in the mood for a *BBQ* today.

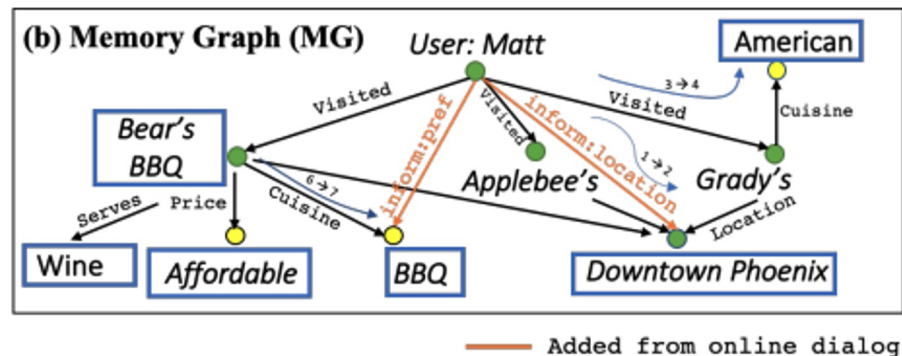
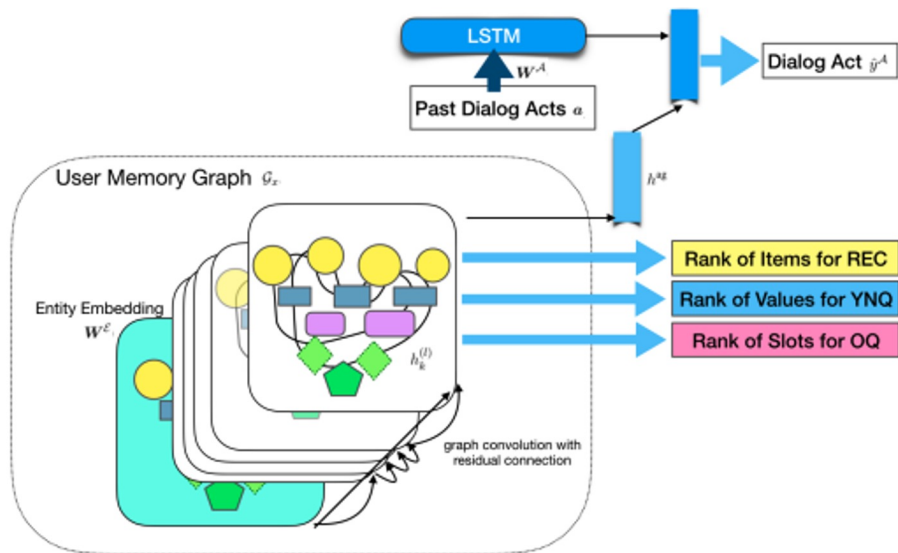
7 *Bear's BBQ* got some *affordable* yet great *BBQ*. They also serve *wine*.

⋮

## (b) Memory Graph (MG)



# Direction 3. Conversational Recommendation



# Direction 3. Conversational Recommendation

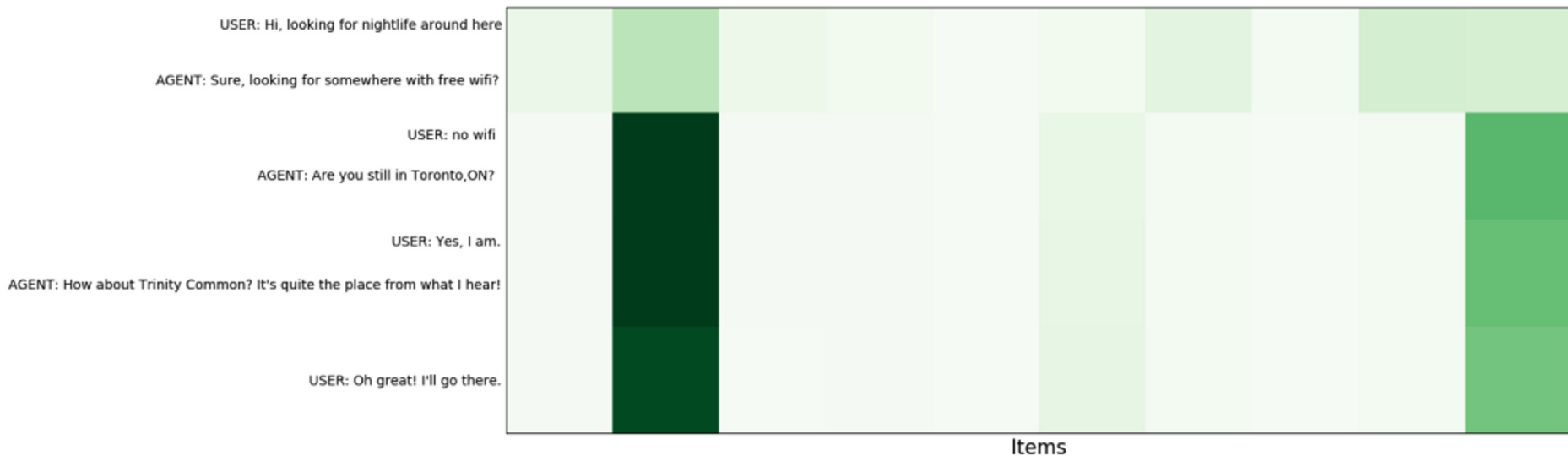
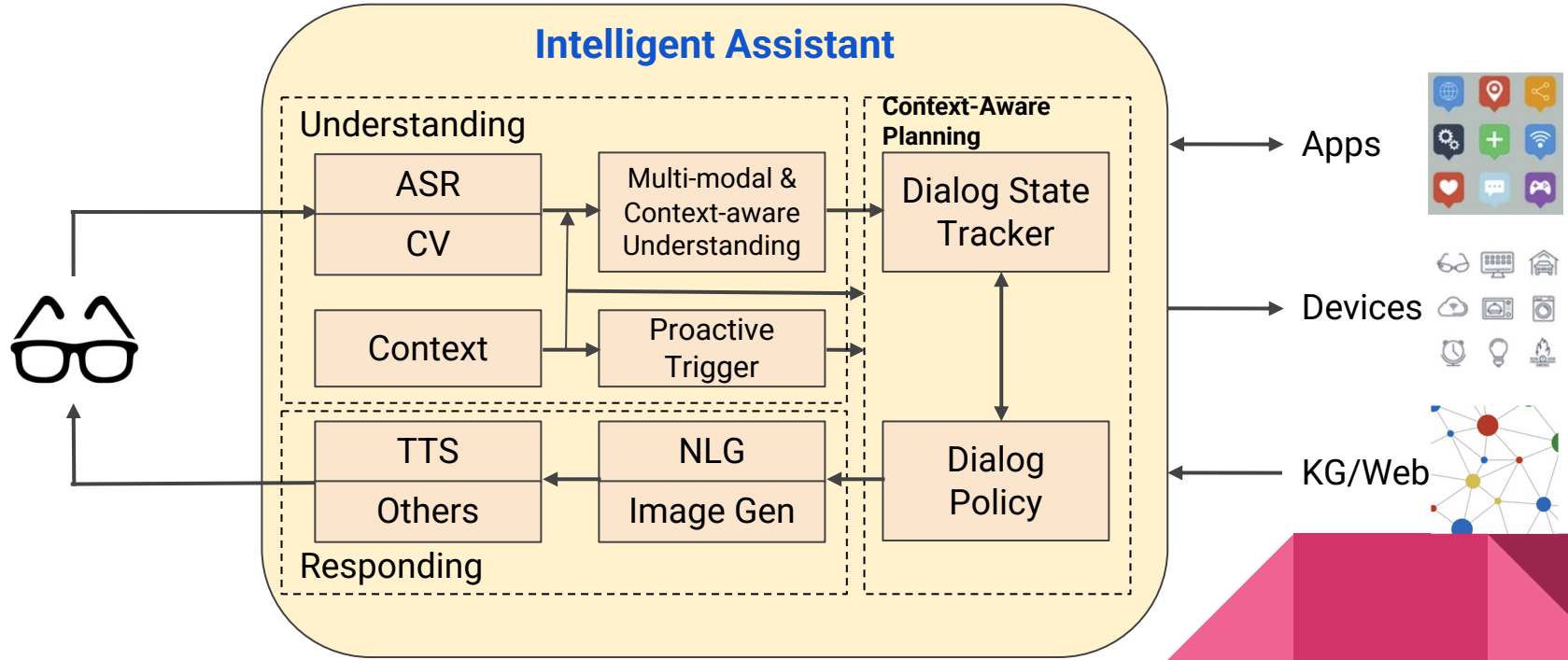


Figure 5: Visualization of item-level conversational reasoning, given an example dialog. Darker color indicates more salient items for recommendation at each given turn (row), predicted by our UMGR model.

# Direction 4. On-Device Machine Learning



**On-device Learning, Federated Learning**

## Direction 4. Federated Learning

- ~~Push data to model~~ → Push models to data

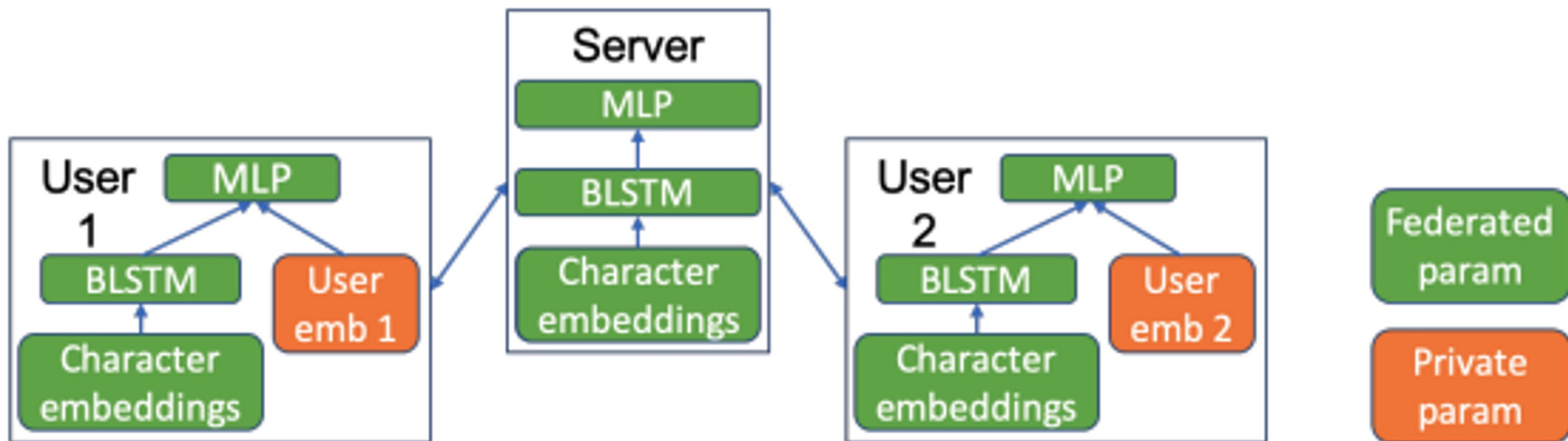


Figure 1: Personalized Document Model in FL.



# Direction 4. On-Device Machine Learning

## Traditional Autoregressive Semantic Parsing

- Pros
  - High accuracy
- Cons: Prohibitively expensive
  - ⇒ Server-side modeling
    - Flaky user experiences w. spotty internet connectivity
    - High latency
    - Compromised user data privacy



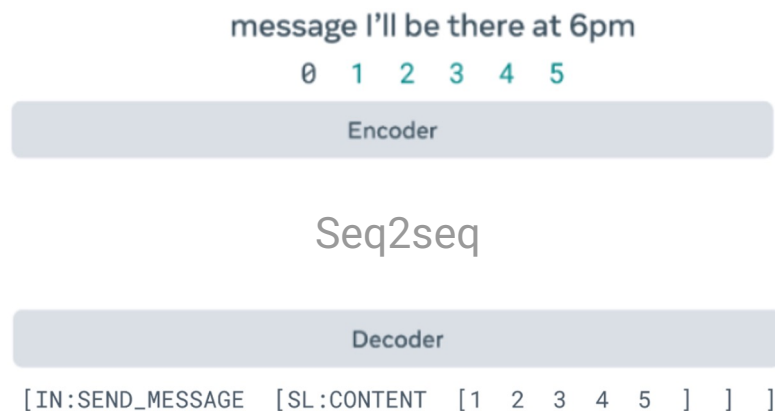
# Direction 4. On-Device Machine Learning

## Non-Autoregressive Semantic Parsing: Parallel prediction



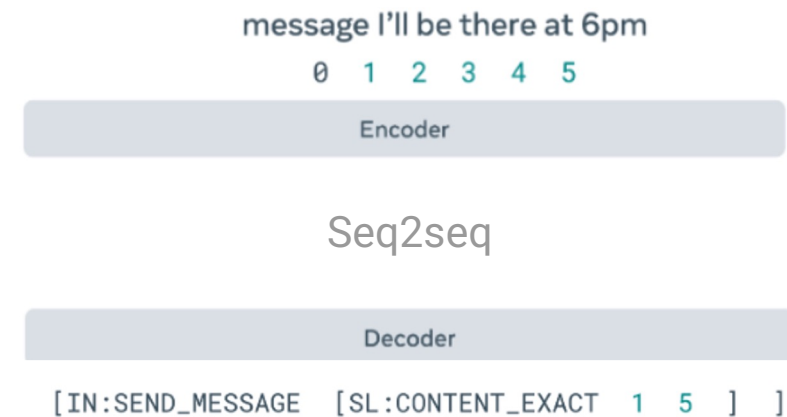
# Direction 4. On-Device Machine Learning

## Non-Autoregressive Semantic Parsing: Parallel prediction

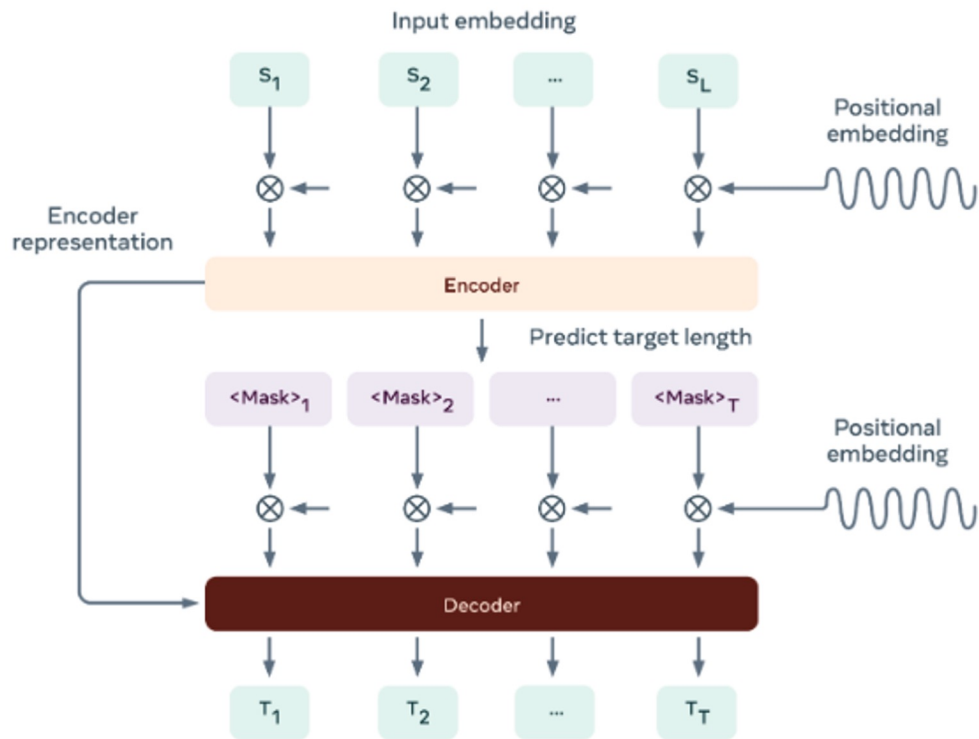


# Direction 4. On-Device Machine Learning

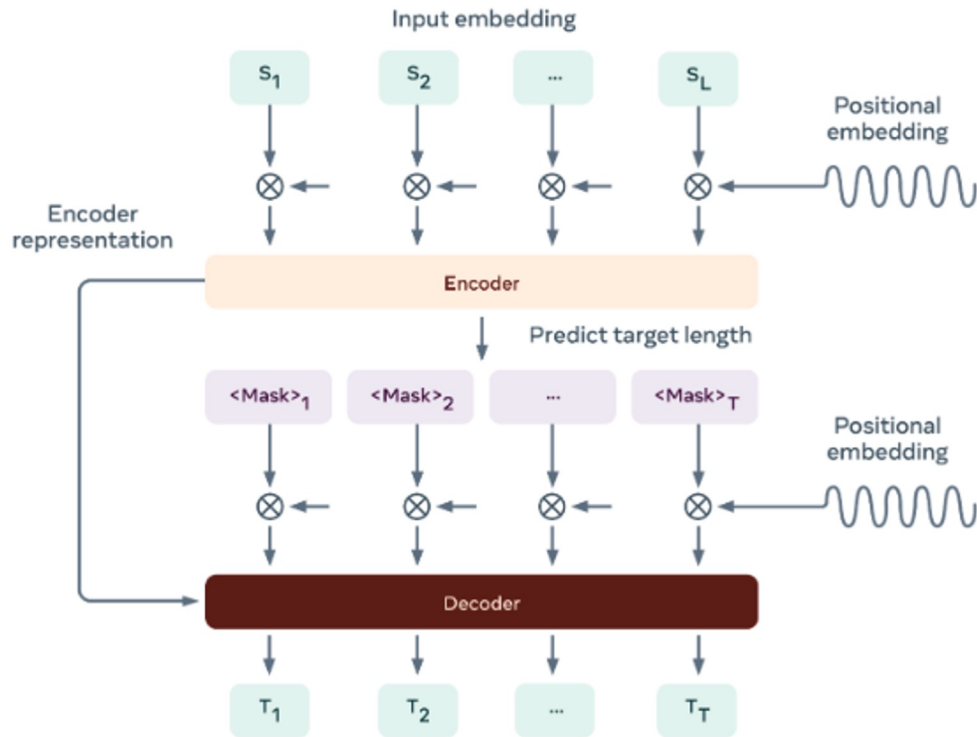
## Non-Autoregressive Semantic Parsing: Parallel prediction



# Direction 4. On-Device Machine Learning

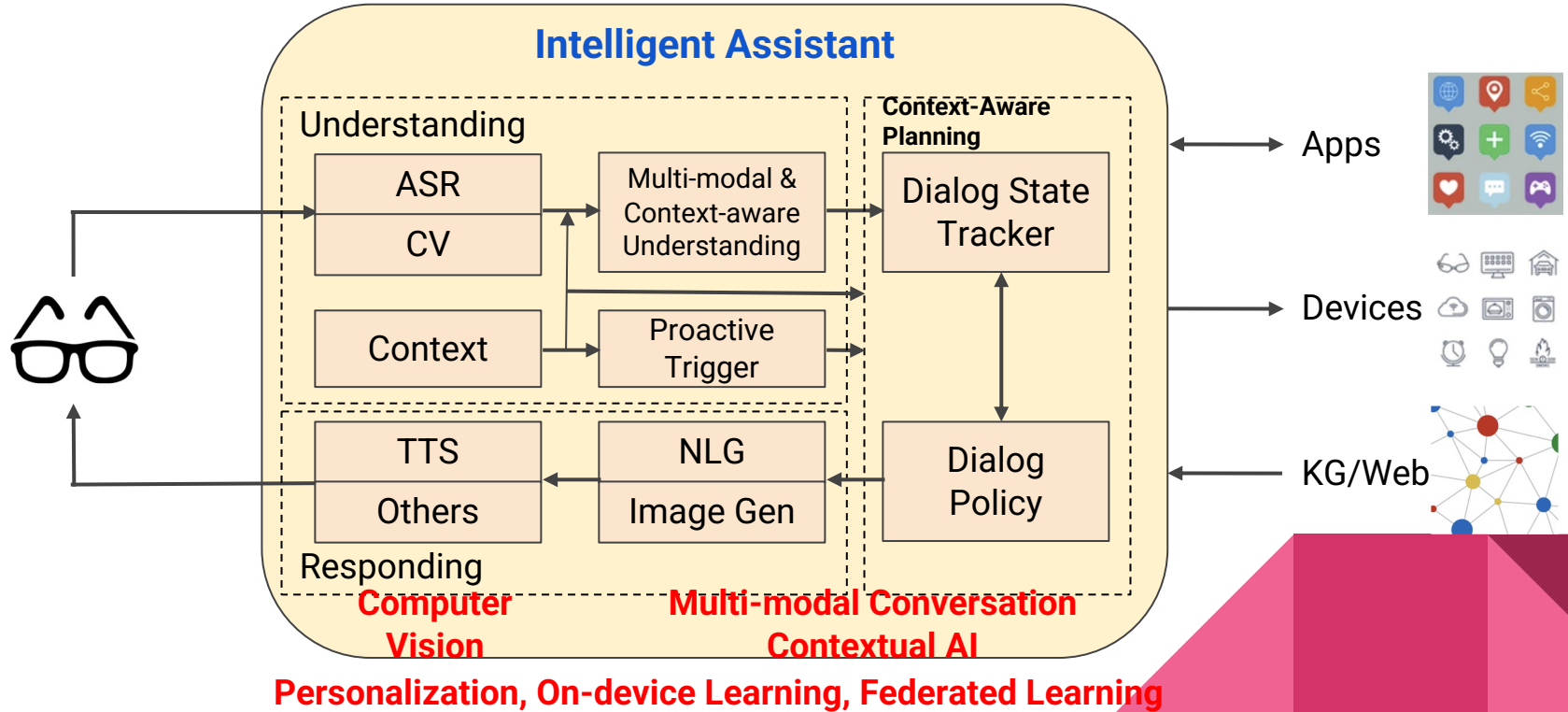


# Direction 4. On-Device Machine Learning



- Memory usage: -83%
- Latency: -70%
- Quality: +1.2% vs. non-AutoRegressive STOA
- Cross-lingual: +14% vs. AutoRegressive baseline

# Recap: New Architecture & Research Areas



# Take-Aways

- An intelligent assistant should be an agent that *knows you and the world*, can *receive your requests* or *predict your needs*, and provide you *the right services at the right time* with your permission
- Next-generation AR/VR assistants require new research on **multi-modal, contextual AI, personalization, on-device learning, federated learning**, etc.
- Multi-modal and context-aware understanding plays a key role in enabling new features







# Thank You

Q&A?