LauzHack

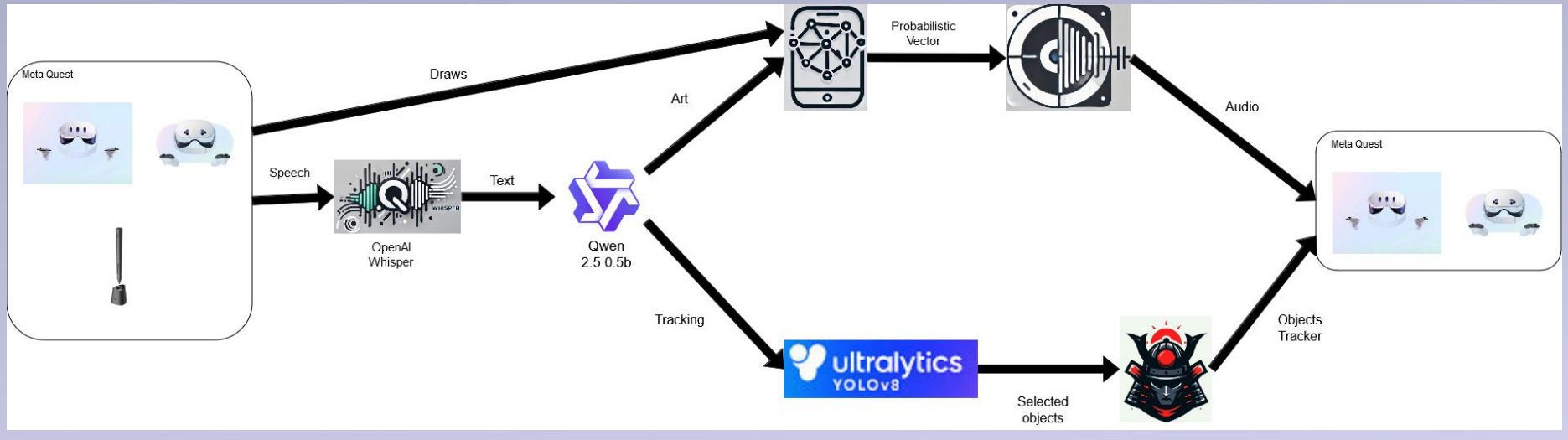
Vitol

# Introduction

- **AI multi-agent** for **identifying and track different objects** in XR multiverse based on **voice commands**
- Combines Speech-To-Text with Qwen[1] + YOLO11 in a **single and efficient architecture**
- Finally, we use a Text-to-Speech to **generate voice responses** to the user

[1] Qwen Technical Report, Alibaba Group, Sep 2023

# Solution - Architecture




**TraceXR architecture**

# Solution - Example

Original image


With objects detected

# Use case: Multi-object track

- **Step 1: Listen to user** speech and convert it to text.
- **Step 2:** The text is analyzed to **understand the user's instructions** using Qwen.
- **Step 3:** The system processes the visual data to **detect all objects** using YOLOv8.
- **Step 4: Hysteresis thresholding** of objects.
- **Step 5: Tracking selected objects** using SAM-Track model (or any other similar models)
- **Step 6: Generate a response** and convert it to speech

# Use case: Object analyzing

- **Step 1: Listen to user** speech, convert it to text and get an object selected by the user.
- **Step 2:** The text is analyzed to **understand the user's instructions** using Qwen and the image is preprocessed (crop, resized, etc…)
- **Step 3:** Using YOLO11 to **detect all objects that satisfy the constraints** from the user
- **Step 4: Compute the number of objects** and return it to Qwen to generate a response
- **Step 5: Generate a response** and convert it to speech


TracXR