Gaze2Prompt: Turning Eye-Tracking Data into Visual Prompts for Multimodal LLMs

Jae Young Choi, Seon Gyeom Kim, Jaywoong Jeong, Ryan Rossi, Jihyung Kil, Tak Yeon Lee

KAIST

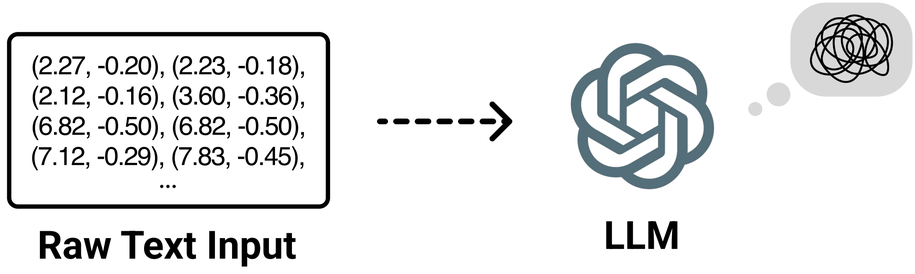
AI EXPERIENCE LAB

Personal Website

Introduction

- The capabilities of **LLMs** are evolving to **comprehend real-world phenomena,** including **human behavior** [1]
- **Sensor data** are commonly used by **LLMs to interpret the physical world,** typically used as **direct text input** [2]

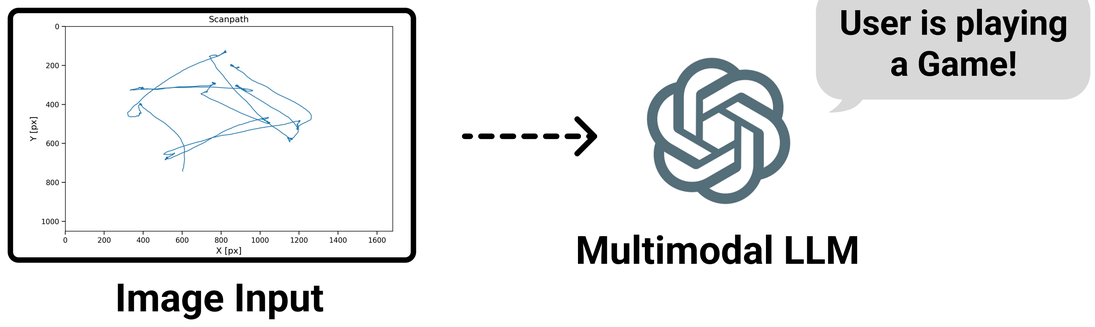
Applying LLMs for Eye-Tracking Data Analysis



Key challenges

1. **Unsuitable** for **numeric data**
2. **Excessive context usage** [3]

Our Approach: Visual Prompting

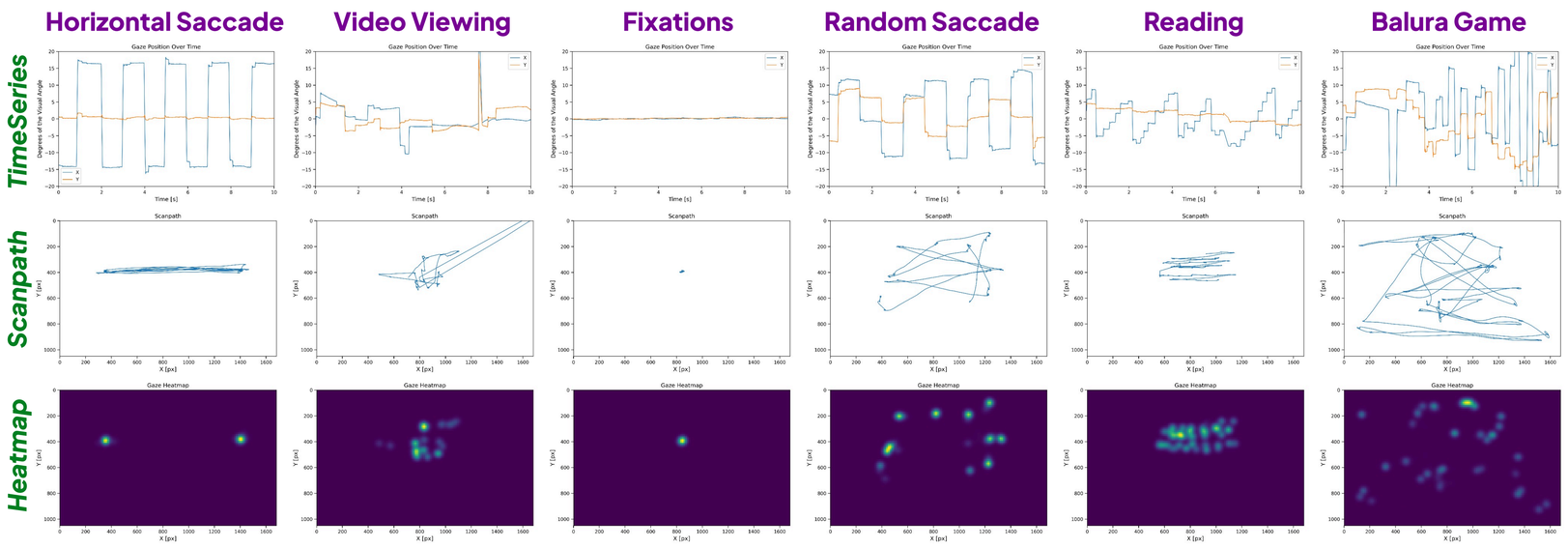


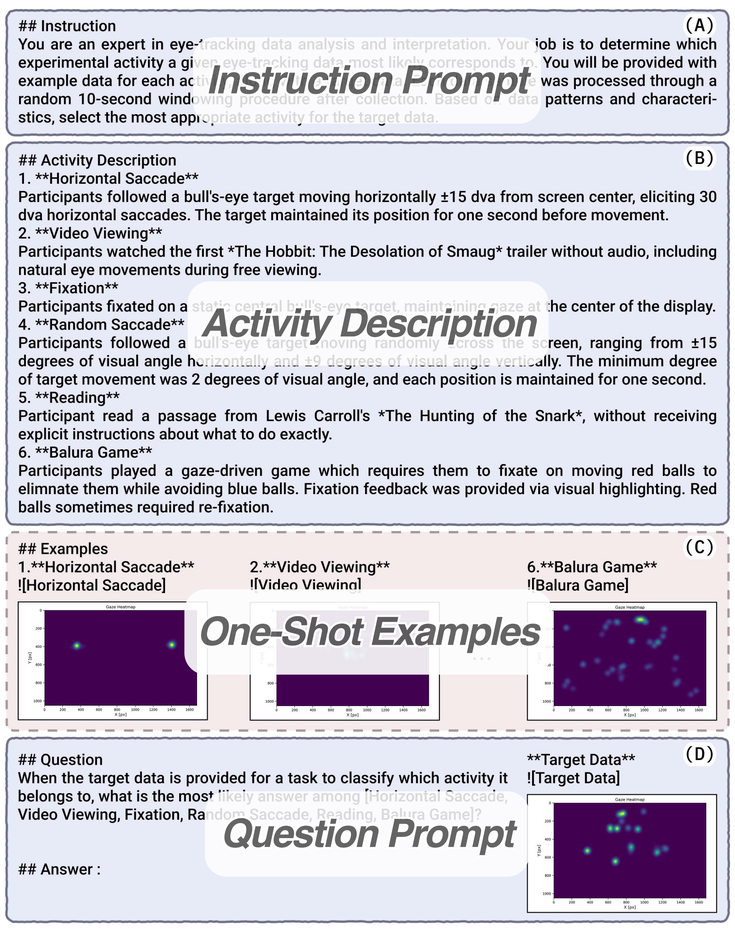
Exploiting **multi-modality** of MLLMs through **visualizations**

Methodology

Human Activity Recognition (HAR) using eye-tracking data

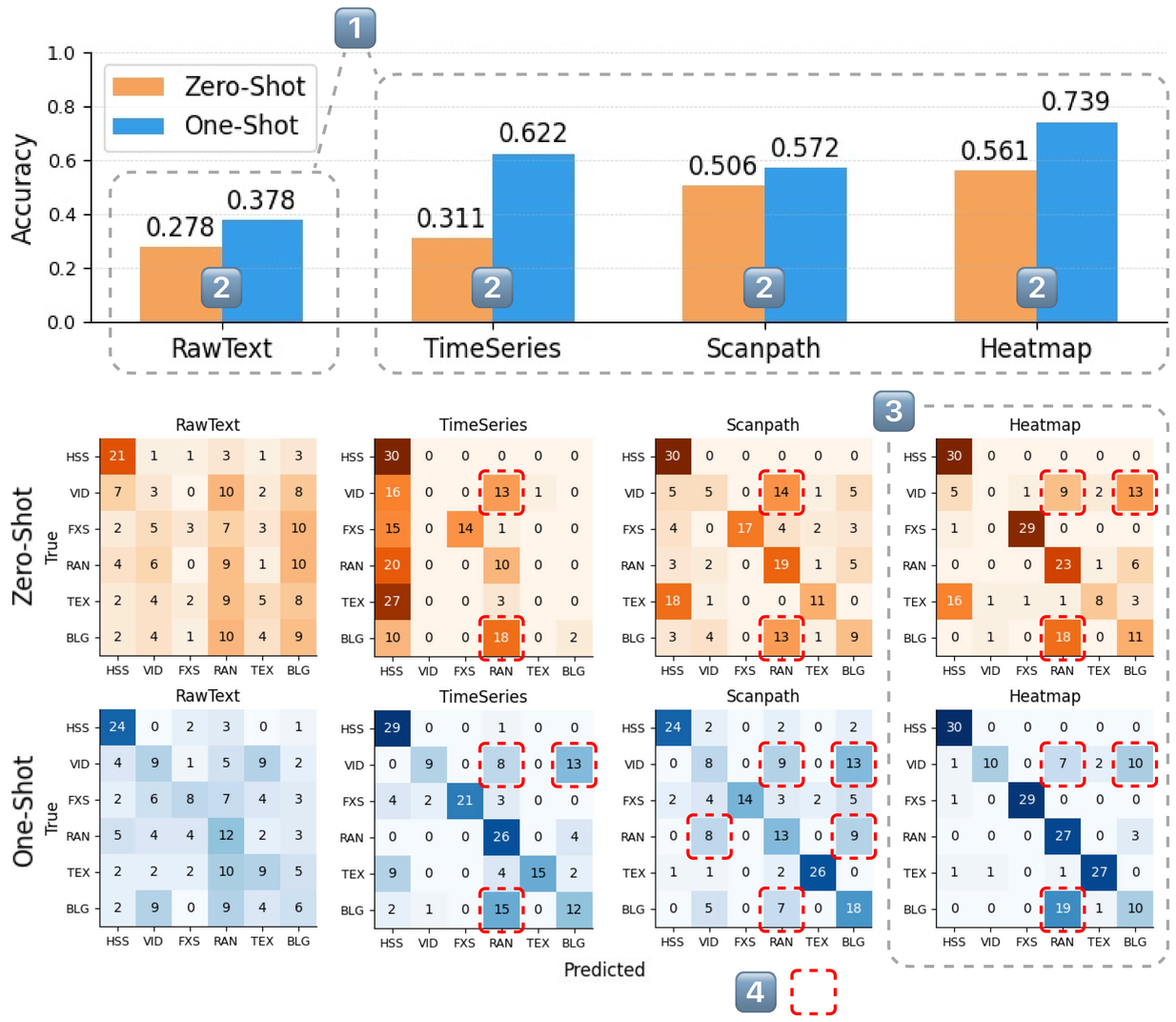
- Dataset: *GazeBase* [4] (322 participants, **6-class activities**)
- Data representations: ***RawText*** *(baseline),* ***TimeSeries, Scanpath, Heatmap*** *(visualizations)*
- **Zero-shot** vs. **One-shot** promping conditions for **GPT-4o** with **structured prompt**





Results

1. **Visual prompt** outperforms **Raw text input**
   - Higher accuracy / Lower token usage (85%↓)
2. **One-shot** outperforms **Zero-shot**
   - though gains vary by visualization and activity
3. **Heatmap is the most effective** overall
   - but scanpath & time-series excel for some tasks (e.g., BLG)
4. **Activity-dependent patterns** are discovered
   - e.g., VID ↔ BLG ↔ RAN show persistent misclassifications



Future Works

- **Enhancing visual representations**
  - e.g., fixation duration visual encoding, stimulus overlay
- **Expanding beyond HAR** to other complex tasks
  - e.g., user identification or cognitive load estimation
- Generalizing approach to **other trajectory data**
  - e.g., movement traces, GPS, robotics

References

[1] Xu, Huatao, et al. "Penetrative ai: Making llms comprehend the physical world." Proceedings of the 25th International Workshop on Mobile Computing Systems and Applications. 2024.
[2] Li, Yuanchun, et al. "Personal llm agents: Insights and survey about the capability, efficiency and security." arXiv preprint arXiv:2401.05459 (2024).
[3] Yoon, Hyungjun, et al. "By My Eyes: Grounding Multimodal Large Language Models with Sensor Data via Visual Prompting." EMNLP. 2024.
[4] Griffith, Henry, et al. "GazeBase, a large-scale, multi-stimulus, longitudinal eye movement dataset." Scientific Data 8.1 (2021): 184.