Tasks Reflected in the Eyes: Egocentric Gaze-Aware Visual Task Type Recognition in Virtual Reality

State Key Laboratory of VR Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing, China

https://zhimin-wang.github.io

Zhimin Wang, Feng Lu





Outline

- Background
- Related Work
- Data Collection
- Our Method
- Experiment Results
- Demo
- Limitations and Future Work



Background

Eye-tracking can be used to recognize users' visual tasks.

user intentions.



(a) The original image

Eye Movement Trajectories During Different Visual Task Executions [Yarbus, 1967]

• Yarbus (1967): Demonstrated that eye movements vary based on tasks, showing that eye-tracking data can reveal

(c) Estimate the circumstances (b) Free examination of the family

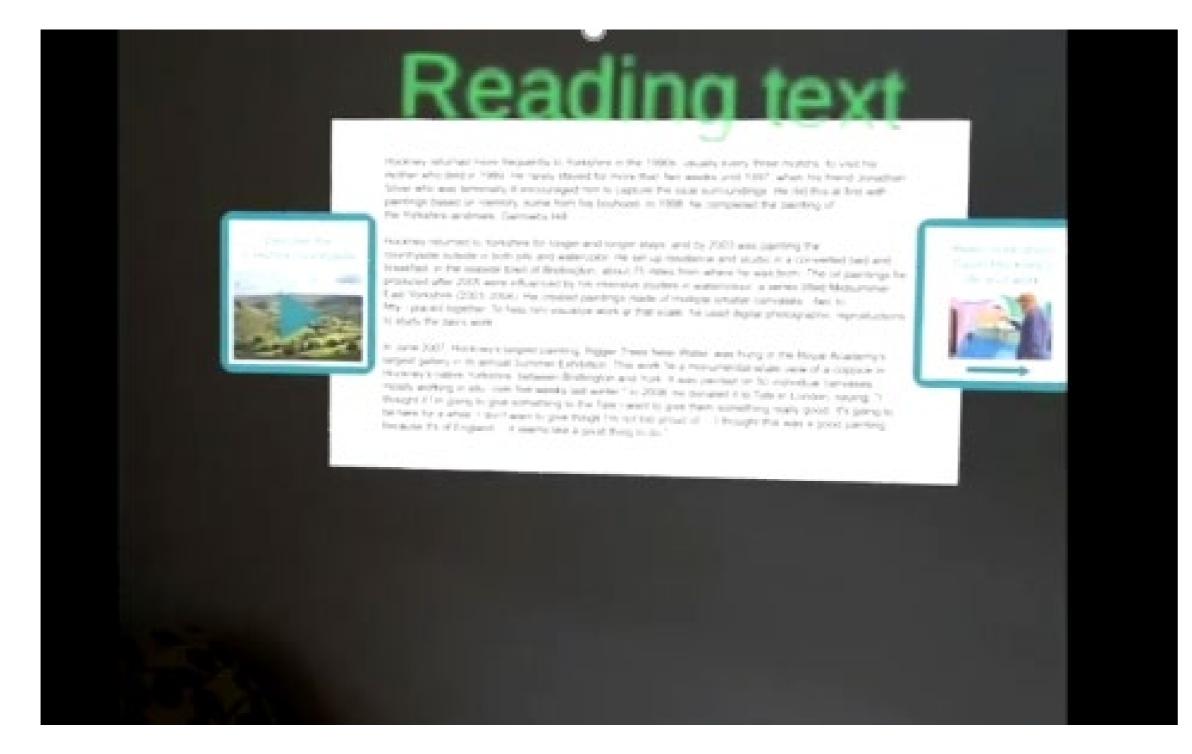




Background

systems.

- immersive atmosphere.



Recognizing visual tasks enables adaptive content design and low-friction interfaces in XR

• For example, when a user is viewing a painting, the XR system could play coordinating music to create an

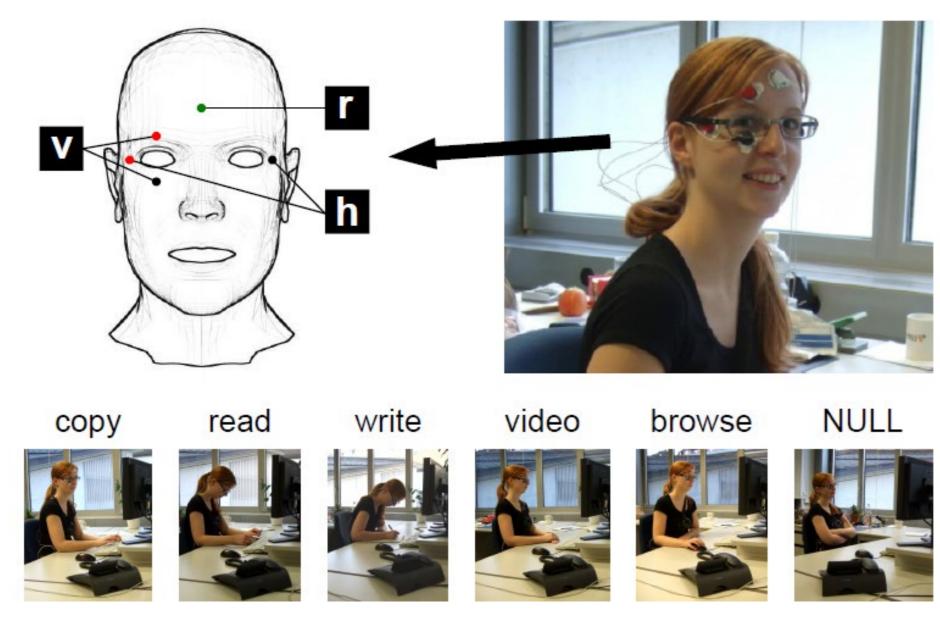
• When the user is reading a text, the XR system can display buttons for page navigation.

Task Recognition-Driven AR Feedback [Lan et al., 2022]

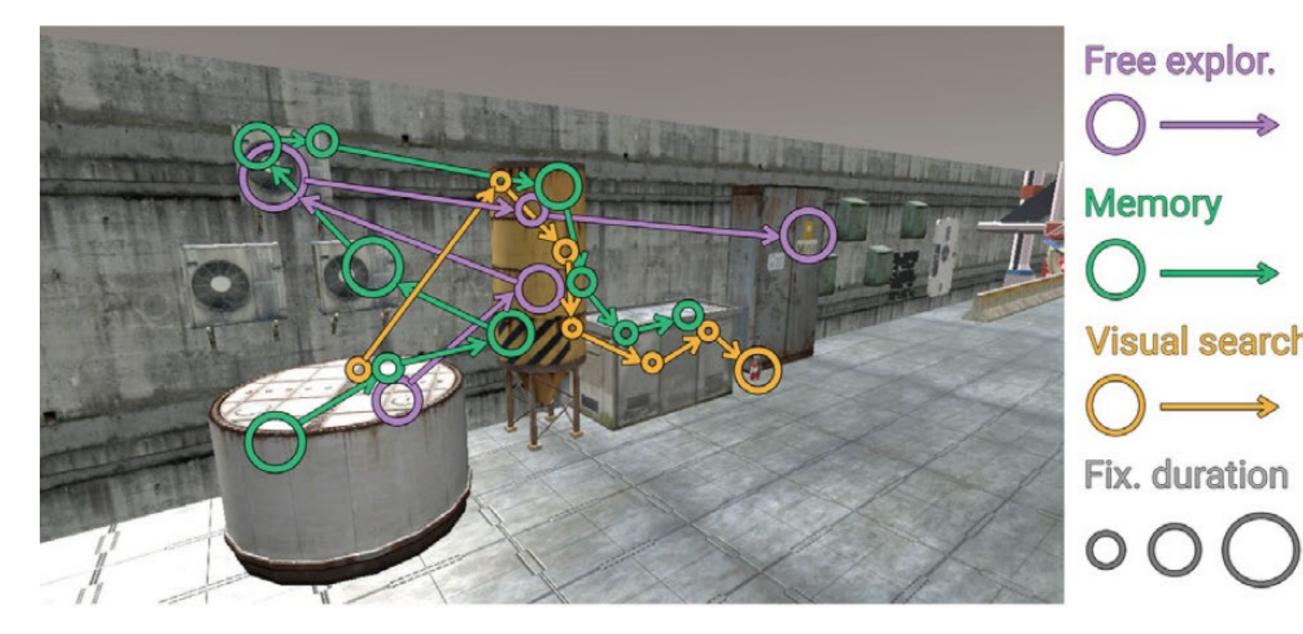


Related Work

- Bulling et al. (2009): Recognized tasks like copying, reading, and web browsing in an office setting using eye-tracking metrics like fixations and saccades.
- Malpica et al. (2023): Recognized tasks, like free exploration, memory, and visual search in indoor corridor using head orientations and gaze directions.



Task Recognition in office setting [Bulling et al., 2009]





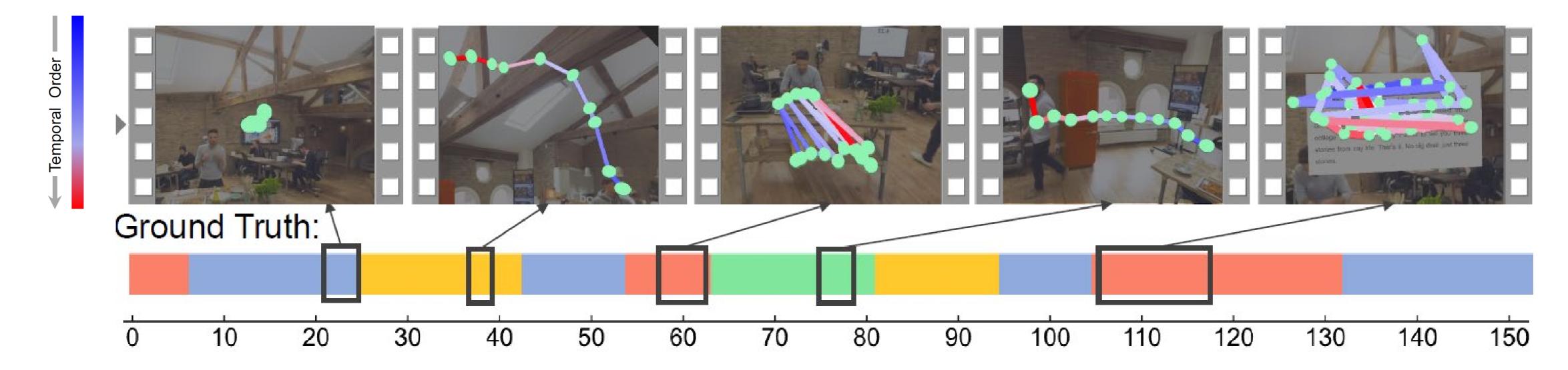
Task Recognition in indoor corridor [Malpica et al., 2023]



Related Work

Challenges:

- VR experiences.



1. Focus on scene-specific tasks, such as tasks designed for office environments or specific

2. No support for task switching: Most studies assume participants perform only one task at a time, which does not reflect real-world scenarios where users often switch between tasks.



Our Contribution

- recognition in a broader range of scenarios.
- recognition method to support free task switching.
- task type recognition through three examples.

• We propose four scene-agnostic visual task types for VR systems, enabling task type

• We provide a new dataset for task type recognition that provides precise temporal boundaries for multiple task types in every video clip, using which we can train the

• We present TRCLP, a novel learning-based approach for recognizing task types, which outperforms the state-of-the-art methods. Additionally, we also demonstrate the utility of



Data Collection – Design of Visual Task Types

Observation:

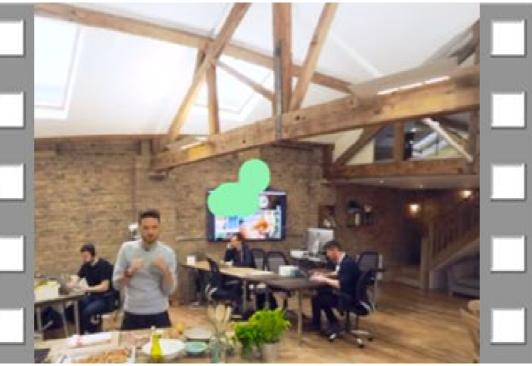
• We noticed **distinct patterns** across various visual tasks.

However, similar tasks, such \bullet as tracking a bicycle rider or tracking a walker, showed similar patterns.





Eye movement patterns differ across various tasks.

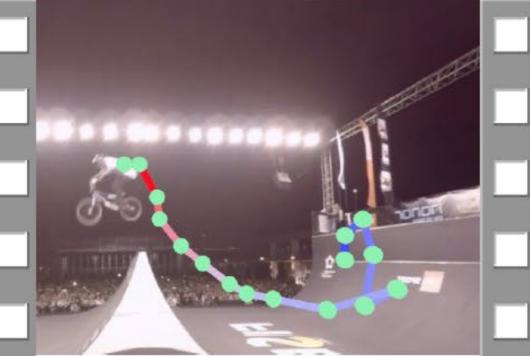


(a) Focus on a TV



(b) Free exploration

Similar tasks tend to produce comparable eye movement patterns.



(a) Track a bicycle rider



(b) Track a walker

(c) Count items

(c) Track a shark

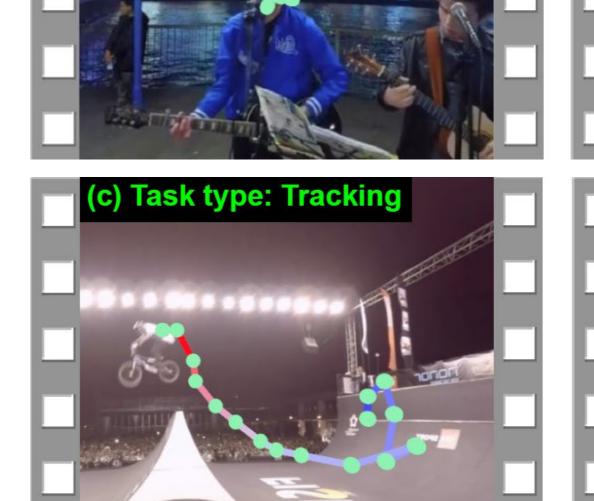
Data Collection – Design of Visual Task Types

Inspired by this observation, we defined four task types based on object states and corresponding eye movement patterns.

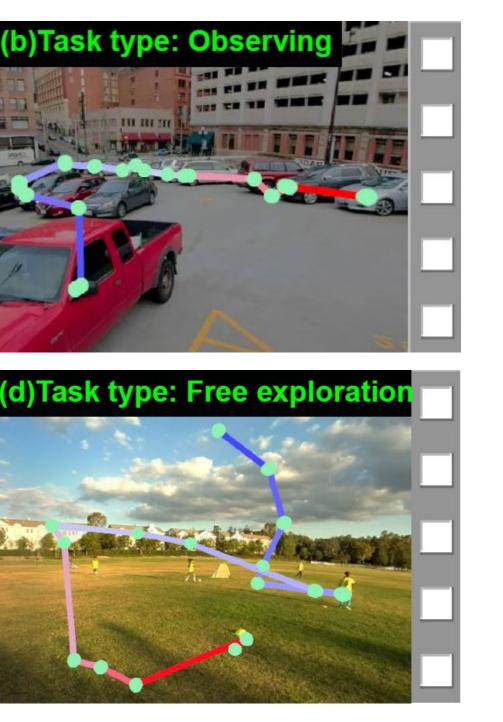
| Visual Task Type | State of Object | Ey |
|---|--|-------|
| Fixating on a Stationary object (FS) | Stationary object | |
| Tracking a Moving object (TM) | Moving object | |
| Observing Sequential objects (OS) | Sequentially stationary objects | Sequ |
| Free Exploration (FE) | Unordered stationary or moving objects | Irreg |

Four visual task types proposed in this study

- e Movement Type
- Long fixations
- Smooth pursuit
- uential saccades with short fixations
- gular saccades with short fixations

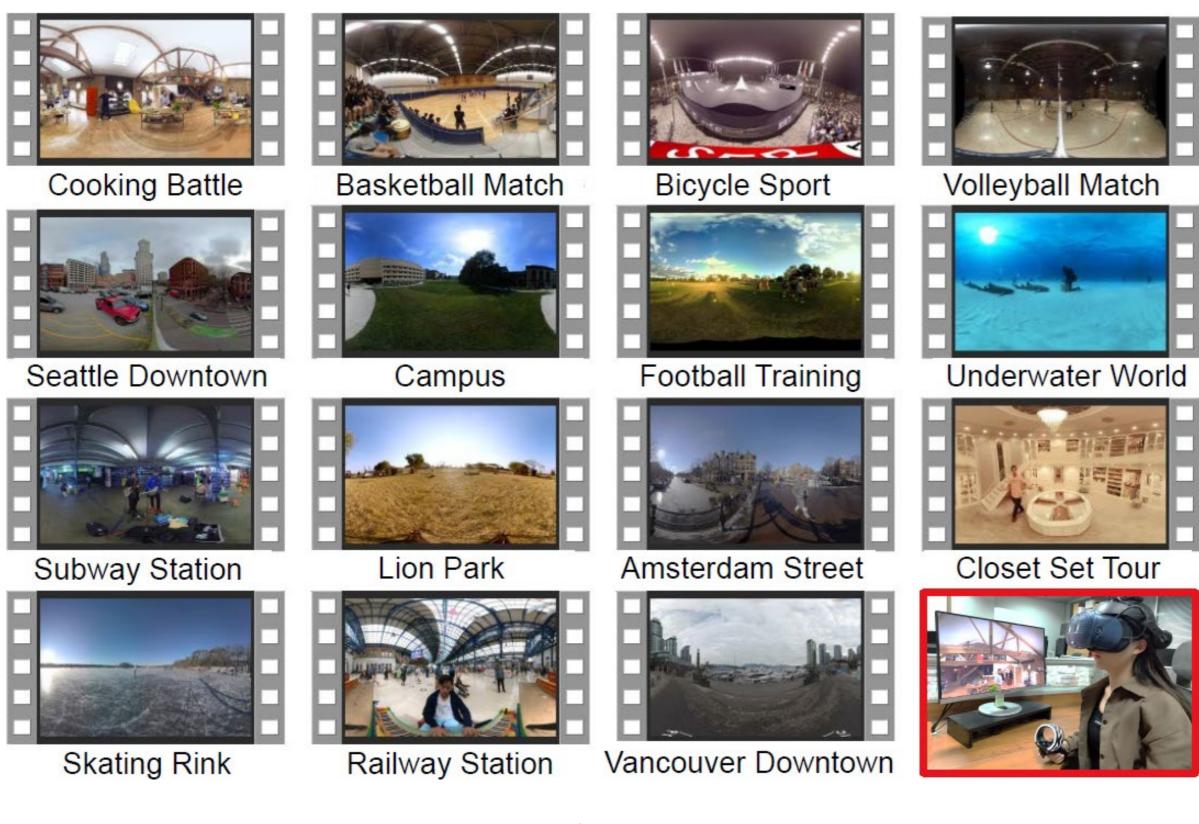


The eye movement patterns related to four task types



Data Collection – Visual Stimuli

Selection of Visual Stimuli: 1) 15 360° VR videos of real-world scenes. 2) 10 text images.



15 real-scene 360° videos

Experimental setup

I had two fathers, a rich one and a poor one. One was highly educated and intelligent; he had a Ph.D. and completed four years of undergraduate work in less than two years. He then went on to Stanford University, the University of Chicago, and Northwestern University to do his advanced studies, all on full financial scholarships. The other father never finished the eighth grade.

1.png

I am honored to be with you today for your commencement from one of the finest universities in the world. Truth be told, I never graduated from college, and this is the closest I've ever gotten to a college graduation. Today I want to tell you three stories from my life. That's it. No big deal. just three stories

4.png

Whenever we read about a scientific breakthrough or a crisis in world affairs, we benefit from the research of those who report it, who in turn benefited from the research of countless others. When we walk into a library, we are surrounded by more than twenty-five centuries of research.

When we go on the Internet, we can read millions of reports written by researchers who have posed questions beyond number gathered untold amounts of information from the research of others to answer them, then shared their answers with the rest of us so that we can carry on their work by asking new questions and, we hope, answering them.

7.png

But even when they agree that writing is an important part of learning, thinking, and understanding, some still wonder why they can't write up their research in their own way, why they have to satisfy demands imposed by a community that they have not joined (or even want to) and conform to conventions they did nothing to create.

10.png

If I had had only one dad, I would have had to accept or reject his advice. Having two dads advising me offered me the choice of contrasting points of view; one of a rich man and one of a poor man.

Instead of simply accepting or rejecting one or the other, I found myself thinking more, comparing and then choosing for myself.

The problem was, the rich man was not rich yet and the poor man not yet poor. Both were just starting out on their careers, and both were strugging with money and families. But they had very different points of view about the subject of money

2.png

And here I was spending all of the money my parents had saved their entire life. So I decided to drop out and trust that it would all work out OK. It was pretty scary at the time, but looking back it was one of the best decisions I ever made. The minute I dropped out I could stop taking the required classes that didn't interest me, and begin dropping in on the ones that looked interesting.

5.png

During the next five years, I started a company named NeXT, another company named Pixar, and fell in love with an amazing woman who would become my wife. Pixar went on to create the worlds first computer animated feature film, Toy Story, and is now the most successful animation studio in the world.

8.png

10 text images for reading

Lesson One: The Rich Don't Work For Money

"Dad, Can You Tell Me How to Get Rich?"

My dad put down the evening paper. "Why do you want to get rich, son?" "Because today Jimmy's mom drove up in their new Cadillac, and they were going to their beach house for the weekend. He took three of his friends, but Mike and I weren't invited. They told us we weren't invited because we were 'poor kids'.'

"They did?" my dad asked incredulously.

"Yeah, they did." I replied in a hurt tone.

My dad silently shook his head, pushed his glasses up the bridge of his nose and went back to reading the paper. I stood waiting for an answer.

3.png

I was lucky. I found what I loved to do early in life. Woz and I started apple in my parents garage when I was 20. We worked hard, and in 10 years apple had grown from just the two of us in a garage into a billion company with over 4000 employees. We just released our finest creation - the Macintosh - a year earlier, and I had just turned 30.

6.png

No one wants to die. Even people who want to go to heaven don't want to die to get there. And yet death is the destination we all share. No one has ever escaped it. And that is as it should be, because Death is very likely the single best invention of Life. It is Life's change agent. It clears out the old to make way for the new.

Right now the new is you, but someday not too long from now, you will gradually become the old and be cleared away. Sorry to be so dramatic, but it is quite true.

9.png

Data Collection - Temporal annotation

Challenge: Since visual tasks are highly subjective and userdependent, it's challenging to accurately determine tasks and task-switching moments during post-hoc annotation.

Solution:

- **Pre-annotation:** For each task type, all possible time \bullet intervals and task instructions are annotated in advance.
- **User-annotation:** During the data collection phase, tasks \bullet are randomly assigned to users, who then determine when to end the task before being assigned another one randomly.

① Input: 360° VR Video 90 s 30 s 60 s 0 s **②** Step 1: Pre-annotation FS Fixate on the host's face. Fixate on the images on the television screen at W direction. os Count the number of items on the shelves at 30~45° direction. Read text. ТΜ Track the face of the walking contestant. Track the moving host's face. FE Freely explore the scene to familiarize yourself with it. ③ Step 2: User-annotation FE FS ΤМ OS FE OS FS

Temporal annotation of task type datasets



150 s 120 s

- Annotate task instructions, start times and end times
- Fixate on the vegetables on the cutting board at 330°.

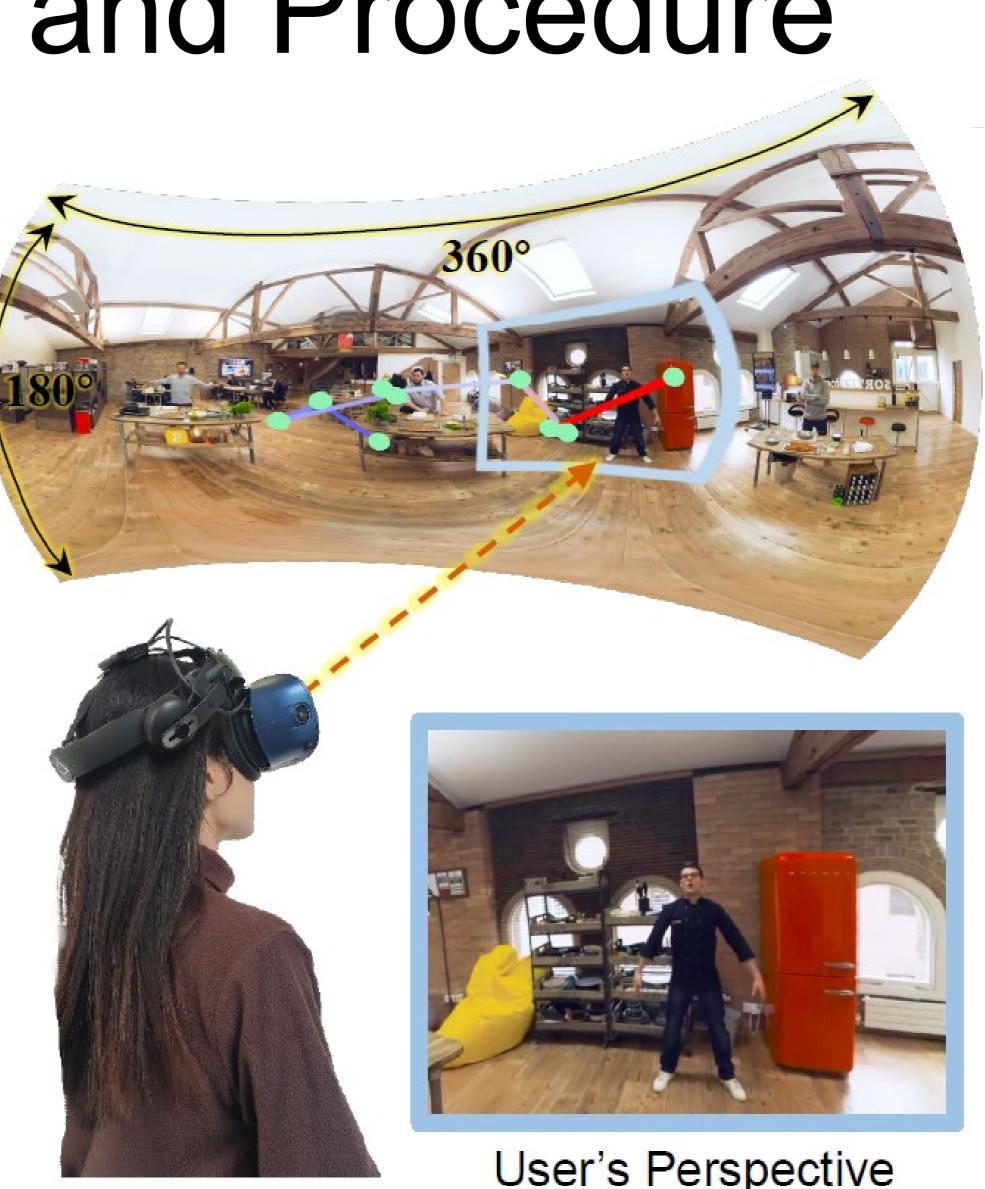
 - Random task types
 - Clear and precise task instructions
 - OS

FE ТΜ

FS

Data Collection - Participants and Procedure

- **Data Recording:** Eye-in-head data, Gaze-in-world data, Head orientation data, Task type labels.
- **Apparatus:** HTC Vive Cosmos.
- **Participants:** 20 subjects (12 male, 8 female).
- **Data Collection Procedure:** Each participant watched 12 randomly selected videos from 15 videos, while performing randomly assigned tasks.
- **Data Volume:** A total of 240 records were obtained (20) participants × 12 videos), with each video being viewed by 16 participants.



Eye movement trajectories and user's perspective



Data Collection - Data Analysis

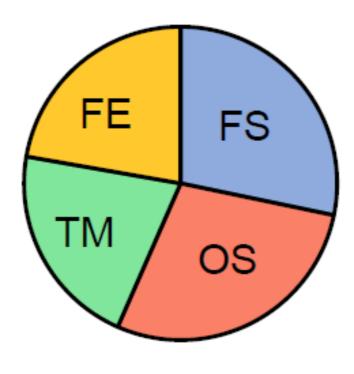
Statistics of the dataset:

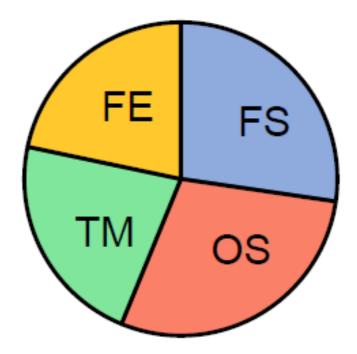
- The total duration of the dataset is 10 hours. \bullet
- The durations of the four task types are relatively balanced (from 21.7% to 28.9%).
- The number of switching between different task types is also relatively balanced.

Statistics of four task types in the dataset

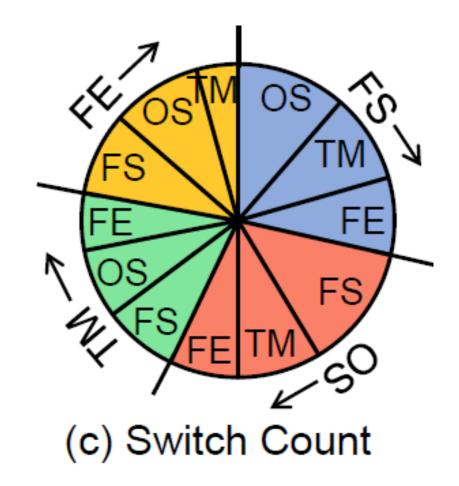
| Task Type Name | FS | OS | TM | FE |
|------------------------------------|-------|-------|-------|-------|
| Task Type Number | 498 | 499 | 373 | 392 |
| Task Type Duration (Total, min) | 163.5 | 173.0 | 132.4 | 130.1 |
| Duration Proportion (%) | 27.3% | 28.9% | 22.1% | 21.7% |







(a) Task Type Number (b) Task Type Duration

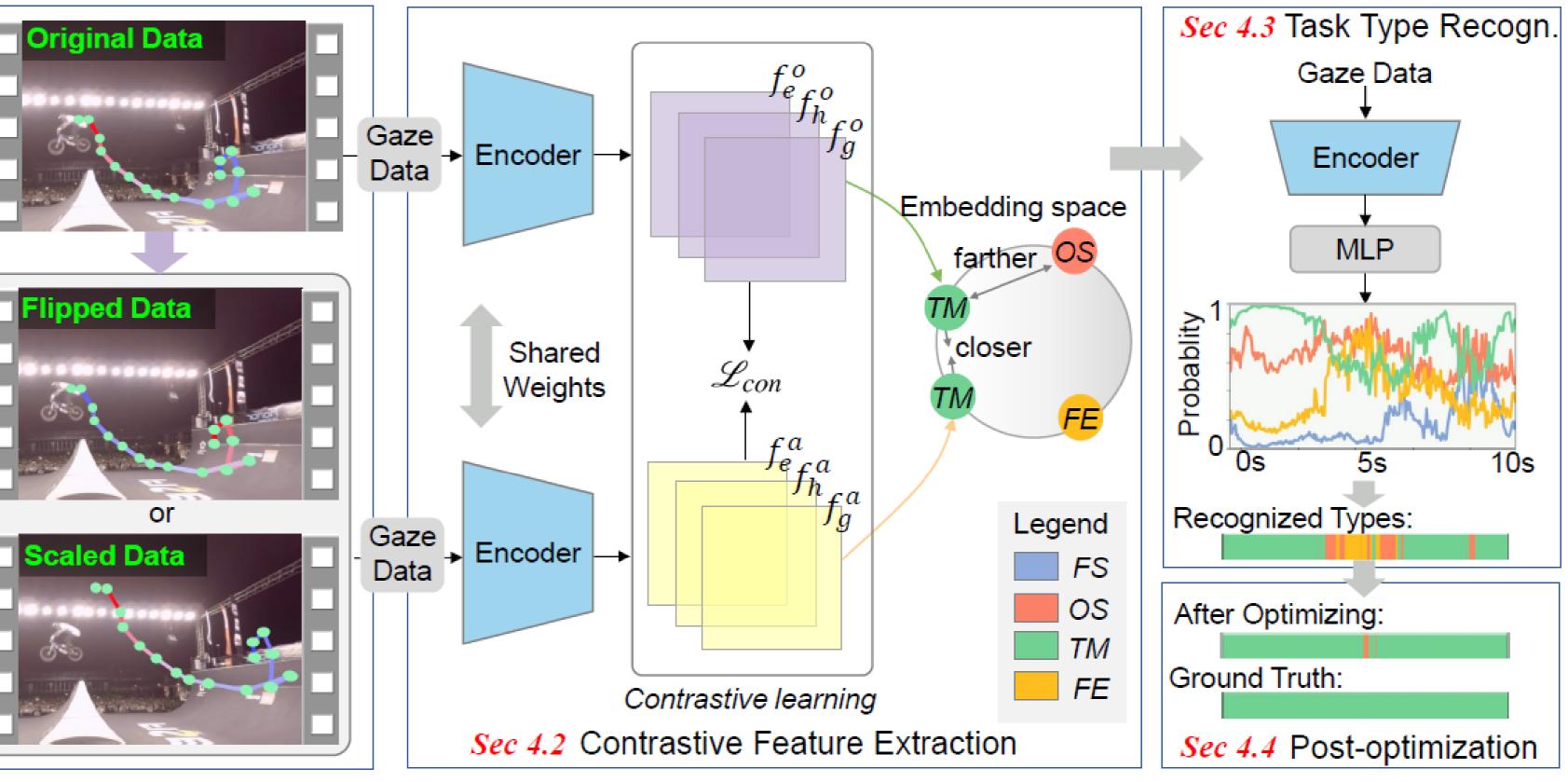


Our Method - Overview

- **Contrastive Feature Extraction:** We apply temporal data augmentation and contrastive learning to extract gaze features.
- Task Type Recognition: A task-type recognition network is designed using CNN, BiGRU, and MLP.
- **Post-Optimization**: A filtering method \bullet is implemented to smooth the recognition results.







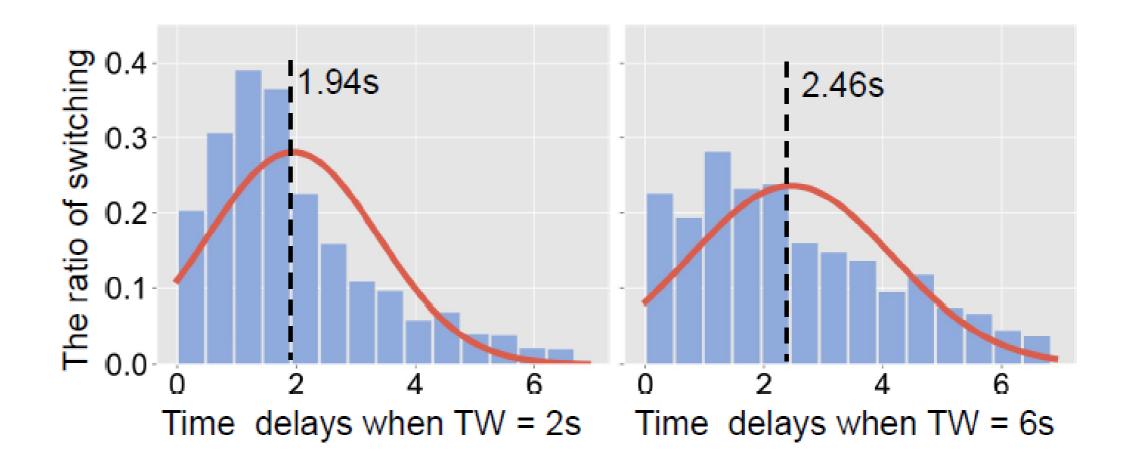
Overview of the proposed TRCLP

Experiment Results

Compare with state-of-the-art methods:

- Unlike other methods that perform best with a 6-second \bullet window, ours excels with a 2-second window, offering the advantage of predicting outcomes 4 seconds earlier.
- With a 2-second input window, our method outperforms lacksquarethe second-best method by 4.3% and 3.5%, respectively.
- The 2-second window data inputs have **shorter time delays** • compared to the 6-second window inputs, offering a 0.5second advantage.

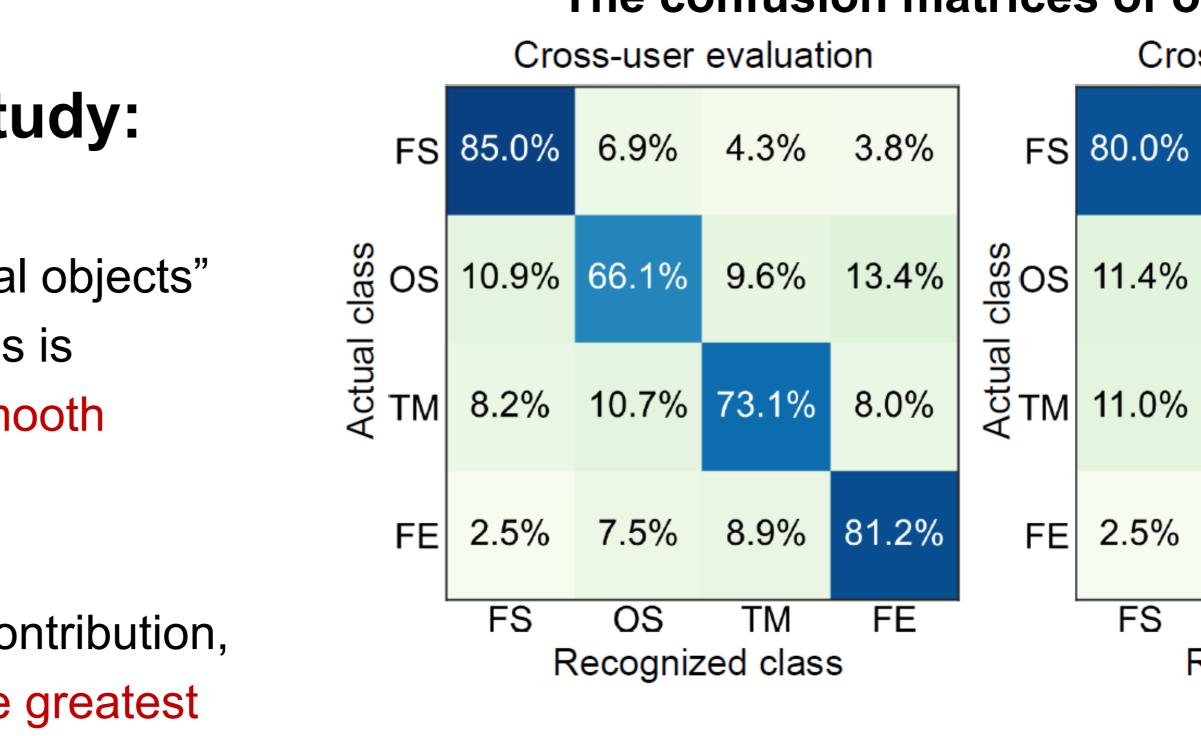
| | Time Window (s) | 2s | <u>6s</u> | 10s |
|-----------------|-----------------|-------|-----------|-------|
| Cross- User | EHTask [21] | 72.1% | 73.3% | 71.0% |
| | RF [20] | 68.3% | 70.1% | 69.4% |
| | MLP-2 [14] | 64.2% | 63.5% | 67.8% |
| | MLP-4 [14] | 65.8% | 70.5% | 70.1% |
| | MLP-6 [14] | 64.9% | 70.2% | 70.0% |
| | CGA [40] | 72.4% | 69.1% | NaN |
| | TRCLP (Ours) | 76.1% | 76.0% | 71.1% |
| Cross- Scene | EHTask [21] | 67.3% | 70.4% | 68.0% |
| | RF [20] | 59.8% | 62.3% | 61.2% |
| | MLP-2 [14] | 60.6% | 63.5% | 62.6% |
| | MLP-4 [14] | 61.3% | 64.3% | 63.7% |
| | MLP-6 [14] | 60.9% | 64.3% | 63.6% |
| | CGA [40] | 68.3% | 65.9% | 64.6% |
| | TRCLP (Ours) | 71.2% | 71.0% | 66.6% |



Experiment Results

Confusion matrices and ablation study:

- Confusion exists between "observing sequential objects" (OS) and "tracking a moving objects" (TM). This is because when object speed exceeds 30°/s, smooth pursuit becomes saccades.
- The ablation study shows each component's contribution, with the post-optimization module providing the greatest improvement.



Ablation study of each component in our method

| Baseline | DA | CL | POST | Cross-User | Cross-Scene |
|--------------|--------------|--------------|--------------|------------|-------------|
| \checkmark | | | | 71.8% | 66.9% |
| \checkmark | \checkmark | | | 72.1% | 67.6% |
| \checkmark | \checkmark | \checkmark | | 72.4% | 67.9% |
| \checkmark | \checkmark | \checkmark | \checkmark | 76.1% | 71.2% |

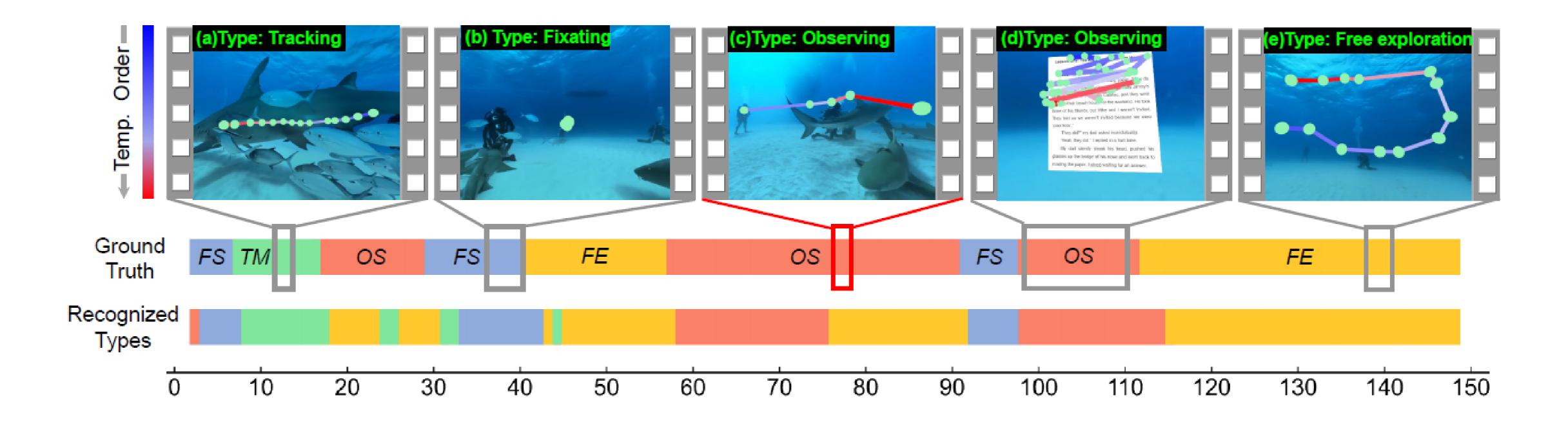
The confusion matrices of our method

Cross-scene evaluation

| 6 | 9.2% | 6.5% | 4.3% |
|----|----------------|-----------------|-------|
| ⁄0 | 61.2% | 13.4% | 14.0% |
| 6 | 16.6% | 59.8% | 12.6% |
|) | 6.5% | 6.1% | 85.0% |
| F | OS Recogniz | TM red class | FE |

Experiment Results

This example shows the eye moveme the VR video of the underwater world.



This example shows the eye movement trajectories of a user while performing four task types in



Limitations and Future Work

- Eye movement overlap: There is confusion between OS and TM tasks when object speed exceeds 30°/s. We plan to improve this by introducing gaze target detection to better differentiate between overlapping eye movements.
- Limitations in scenarios: Our current setup requires participants to remain stationary. • Future work will explore scenarios with user movement to better reflect real-world conditions.
- More complex task types: Irregular saccades is currently represented by FE. This eye movement pattern is complex and can correspond to multiple task types not fully explored in this paper. Future work will explore the recognition of more task types involving irregular saccades.

Tasks Reflected in the Eyes: Egocentric Gaze-Aware Visual Task Type Recognition in Virtual Reality

Our dataset and source code are available at: https://zhimin-wang.github.io

Thank you!



