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

Zhimin Wang, Feng Lu

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

https://zhimin-wang.github.io

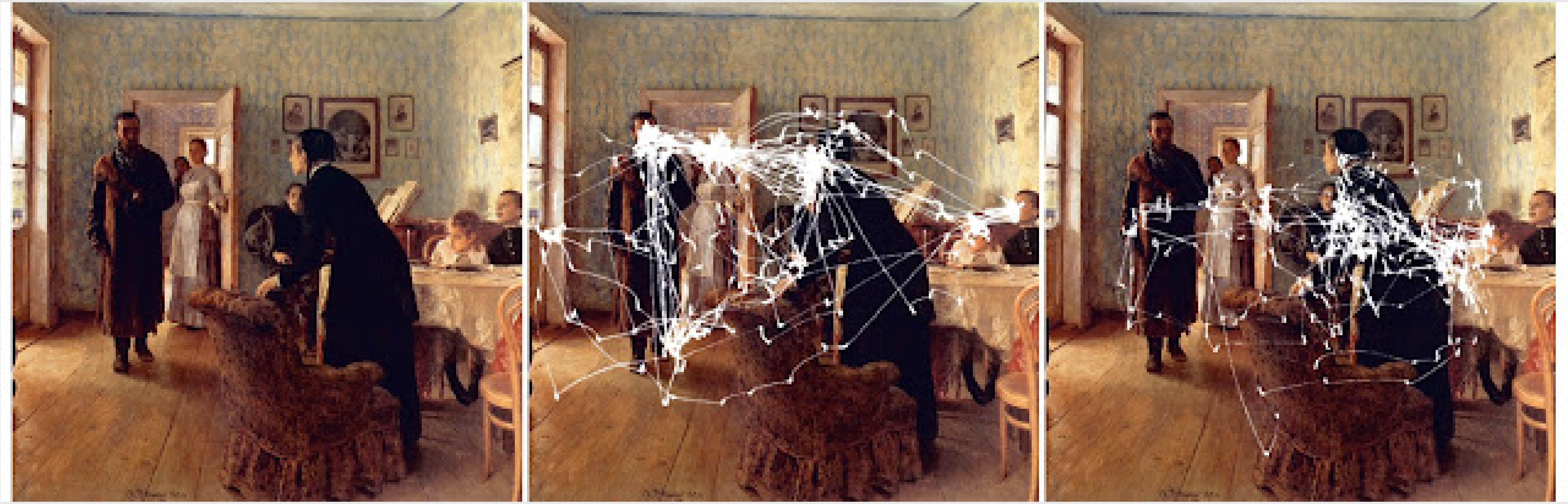
Outline

- 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.

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



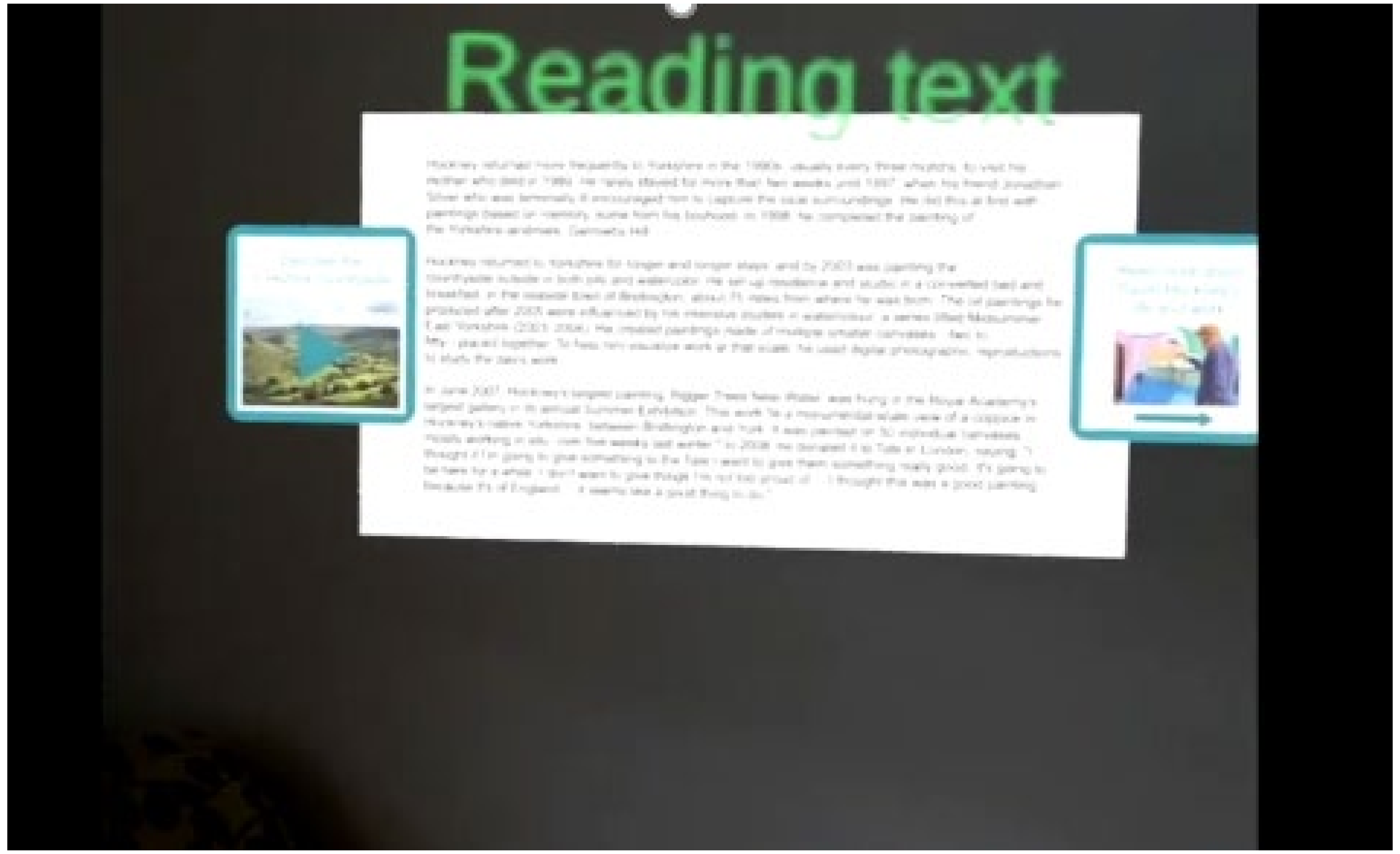
(a) The original image (b) Free examination (c) Estimate the circumstances of the family

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

Background

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

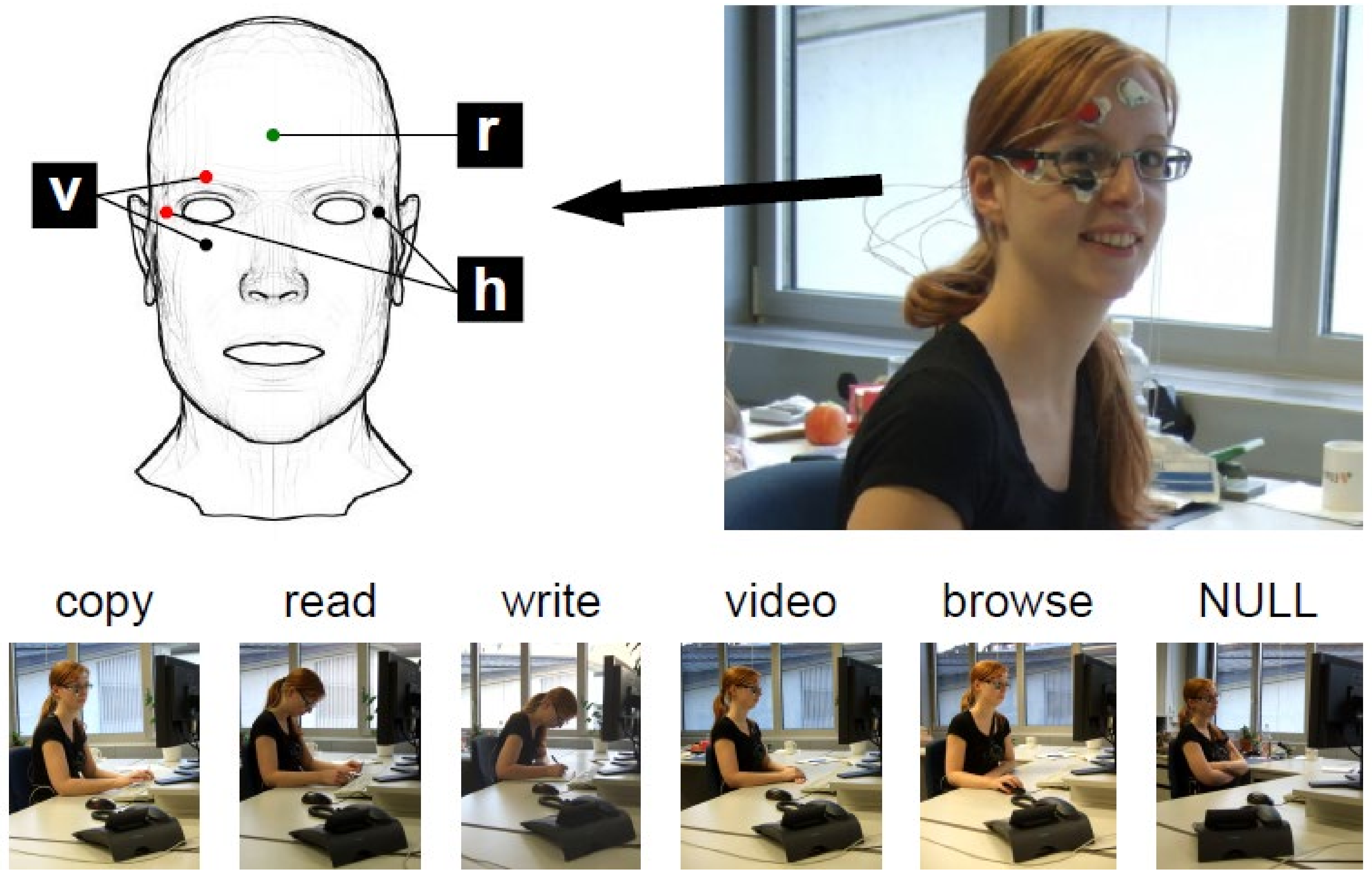
- For example, when a user is viewing a painting, the XR system could play coordinating music to create an immersive atmosphere.
- When the user is reading a text, the XR system can display buttons for page navigation.



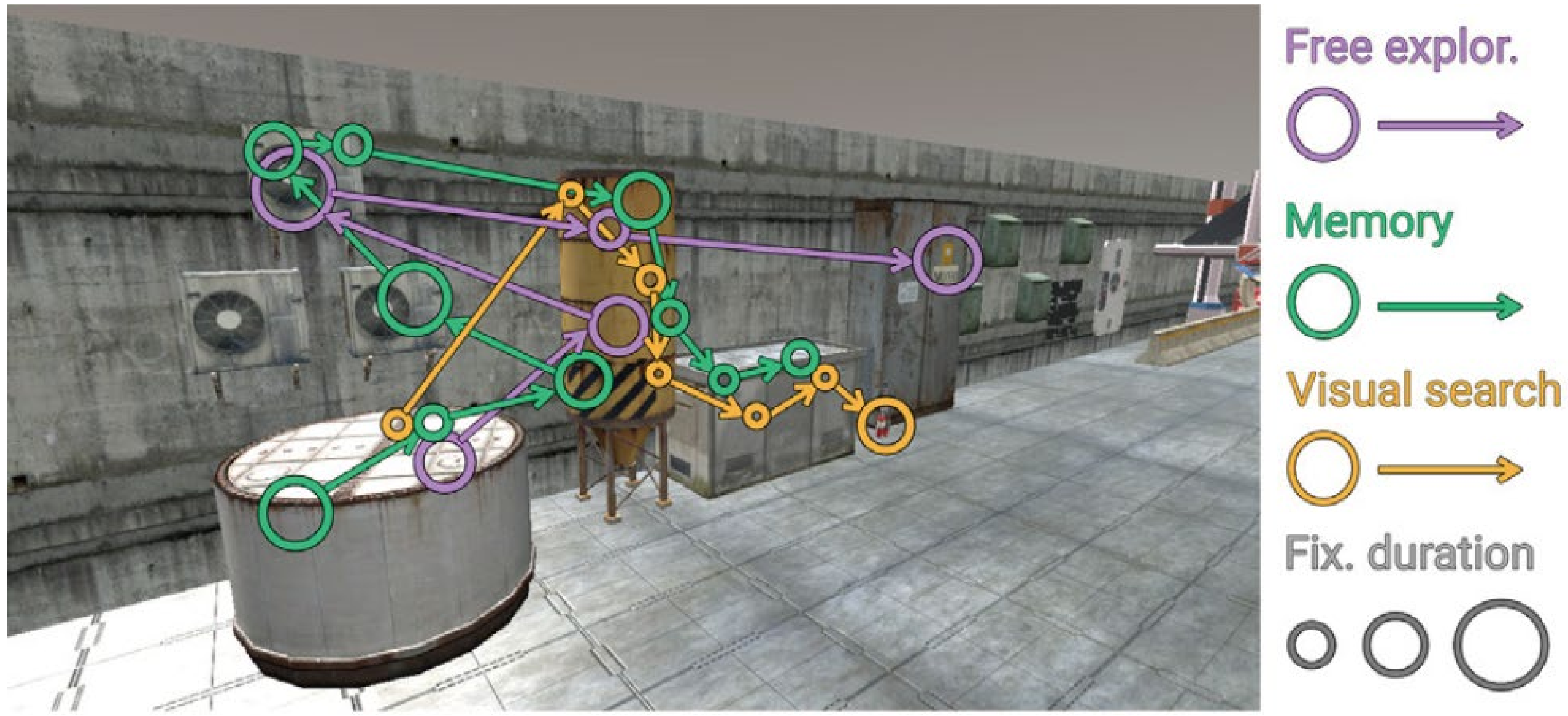
Task Recognition-Driven AR Feedback
[Lan et al., 2022]

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]

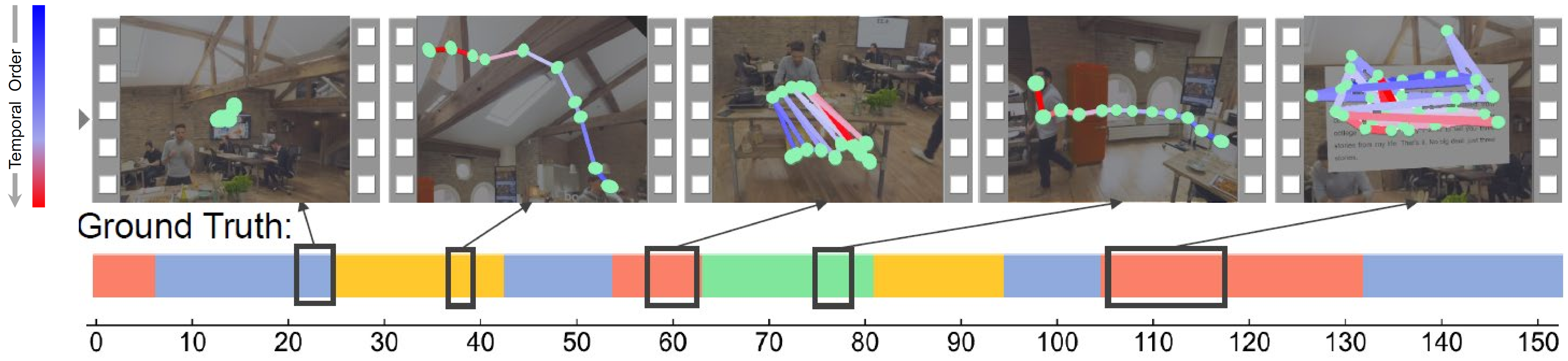


Task Recognition in indoor corridor [Malpica et al., 2023]

Related Work

Challenges:

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

- We propose **four scene-agnostic visual task types** for VR systems, enabling task type recognition in a broader range of scenarios.
- 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 recognition method to support free task switching**.
- 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 task type recognition through **three examples**.

Data Collection – Design of Visual Task Types

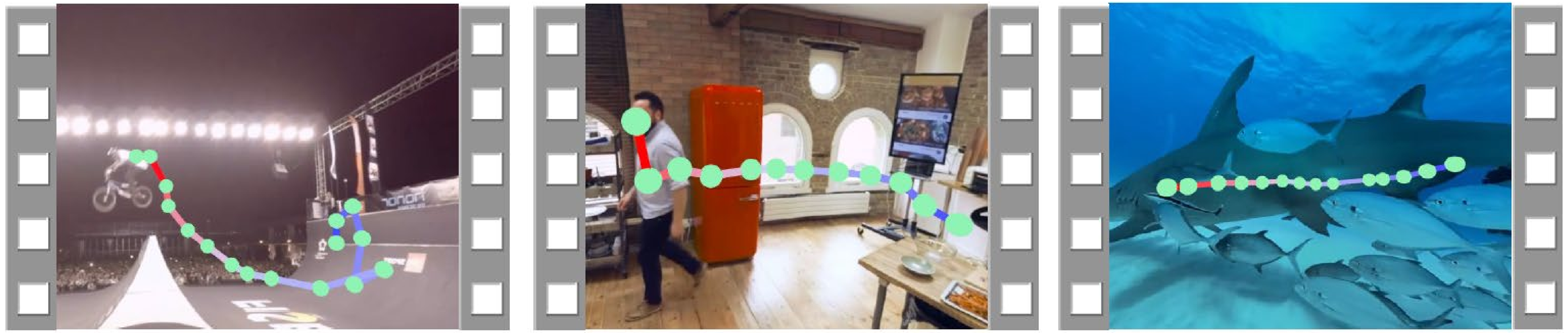
Observation:

- We noticed **distinct patterns** across **various visual tasks**.
- However, **similar tasks**, such 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 (c) Count items

Similar tasks tend to produce comparable eye movement patterns.



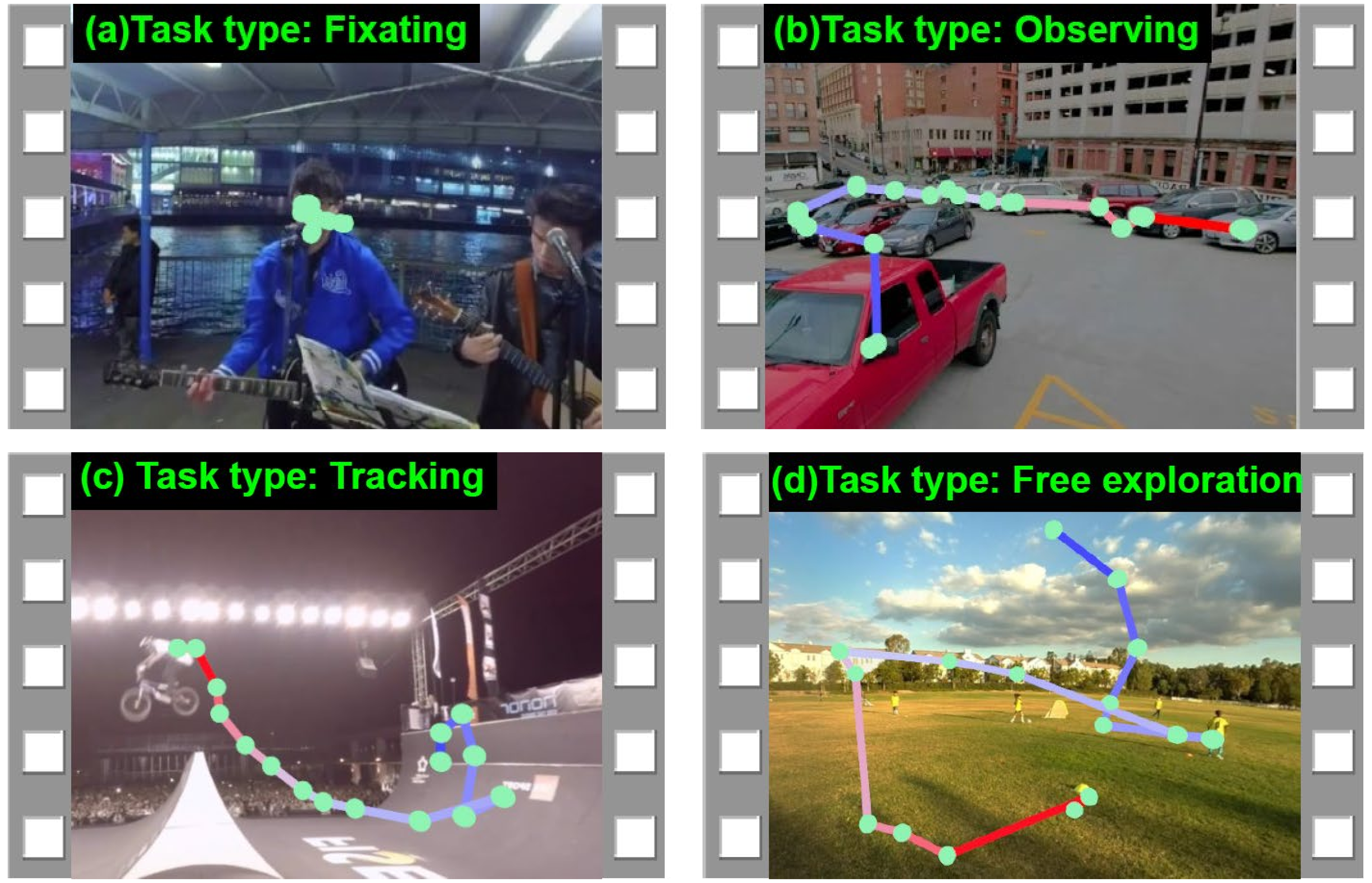
(a) Track a bicycle rider (b) Track a walker (c) Track a shark

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	Eye Movement Type
Fixating on a Stationary object (FS)	Stationary object	Long fixations
Tracking a Moving object (TM)	Moving object	Smooth pursuit
Observing Sequential objects (OS)	Sequentially stationary objects	Sequential saccades with short fixations
Free Exploration (FE)	Unordered stationary or moving objects	Irregular saccades with short fixations

Four visual task types proposed in this study

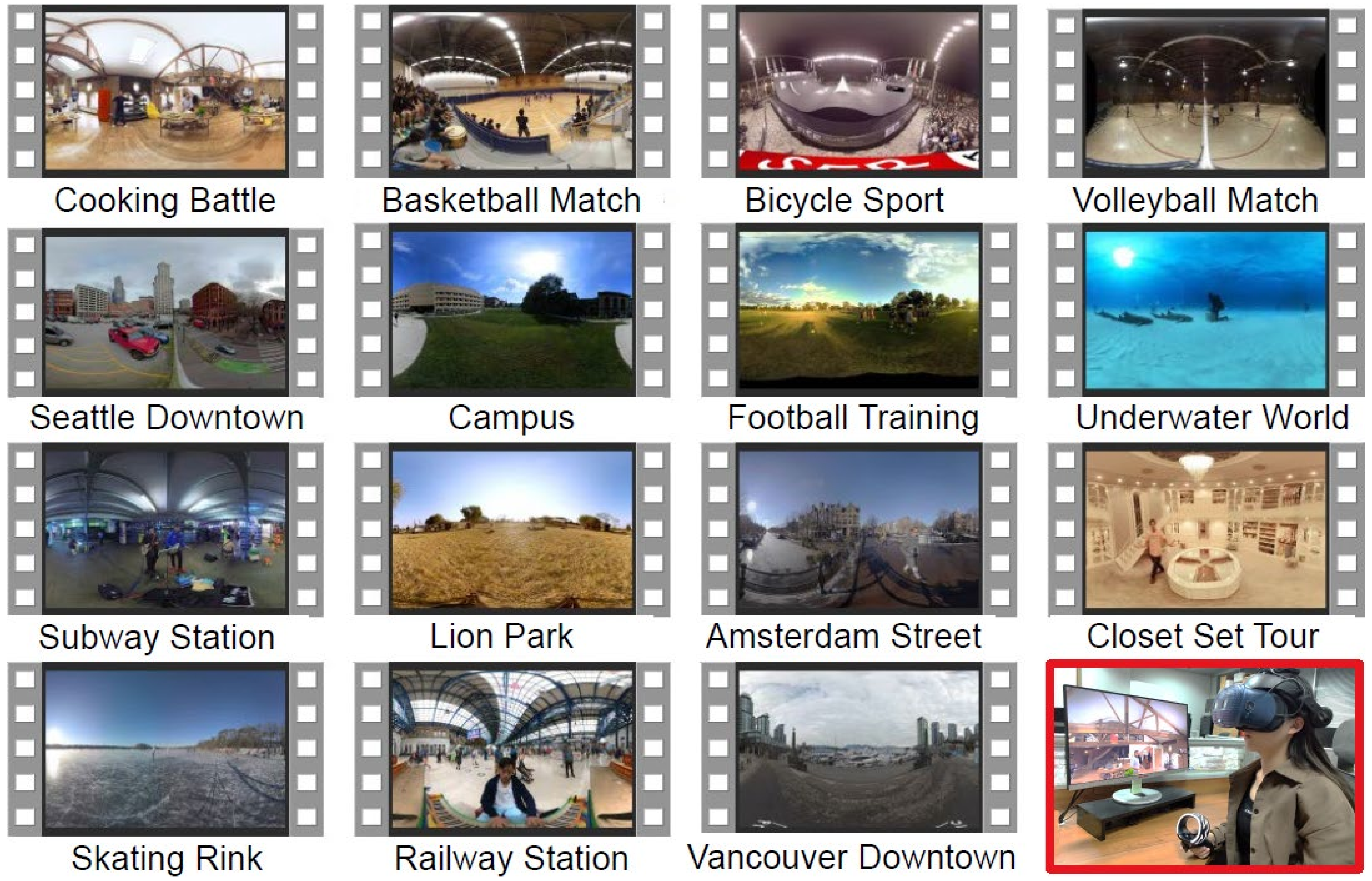


The eye movement patterns related to four task types

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**

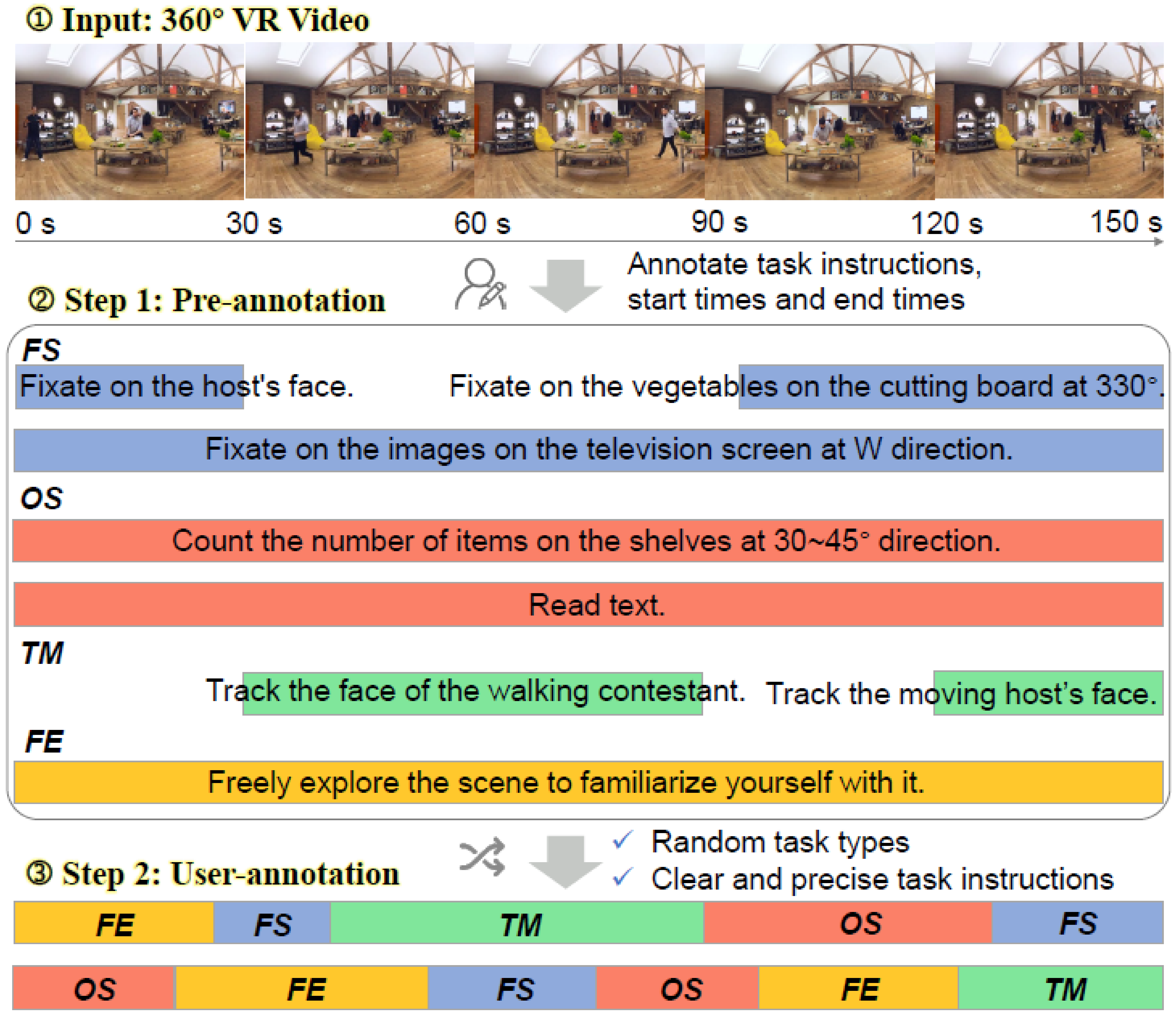
10 text images for reading

Data Collection - Temporal annotation

Challenge: Since visual tasks are highly subjective and user-dependent, 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 intervals and task instructions are annotated in advance.
- **User-annotation:** During the data collection phase, tasks are randomly assigned to users, who then determine when to end the task before being assigned another one randomly.



Temporal annotation of task type datasets

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

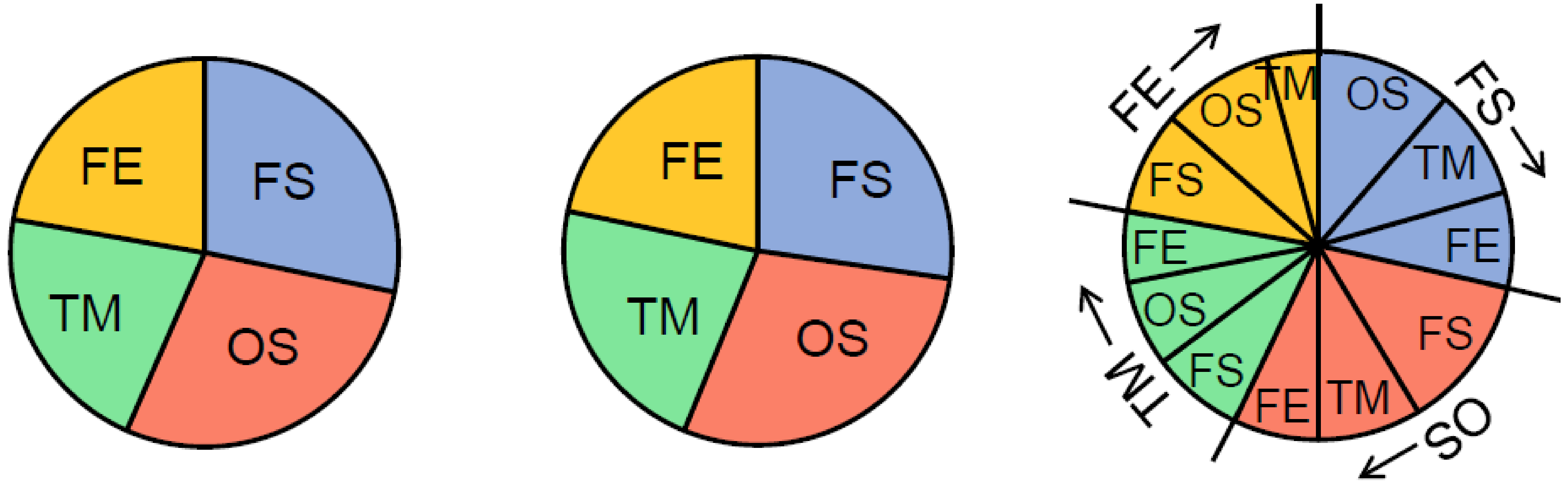
Data Collection - Data Analysis

Statistics of the dataset:

- The total duration of the dataset is 10 hours.
- 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

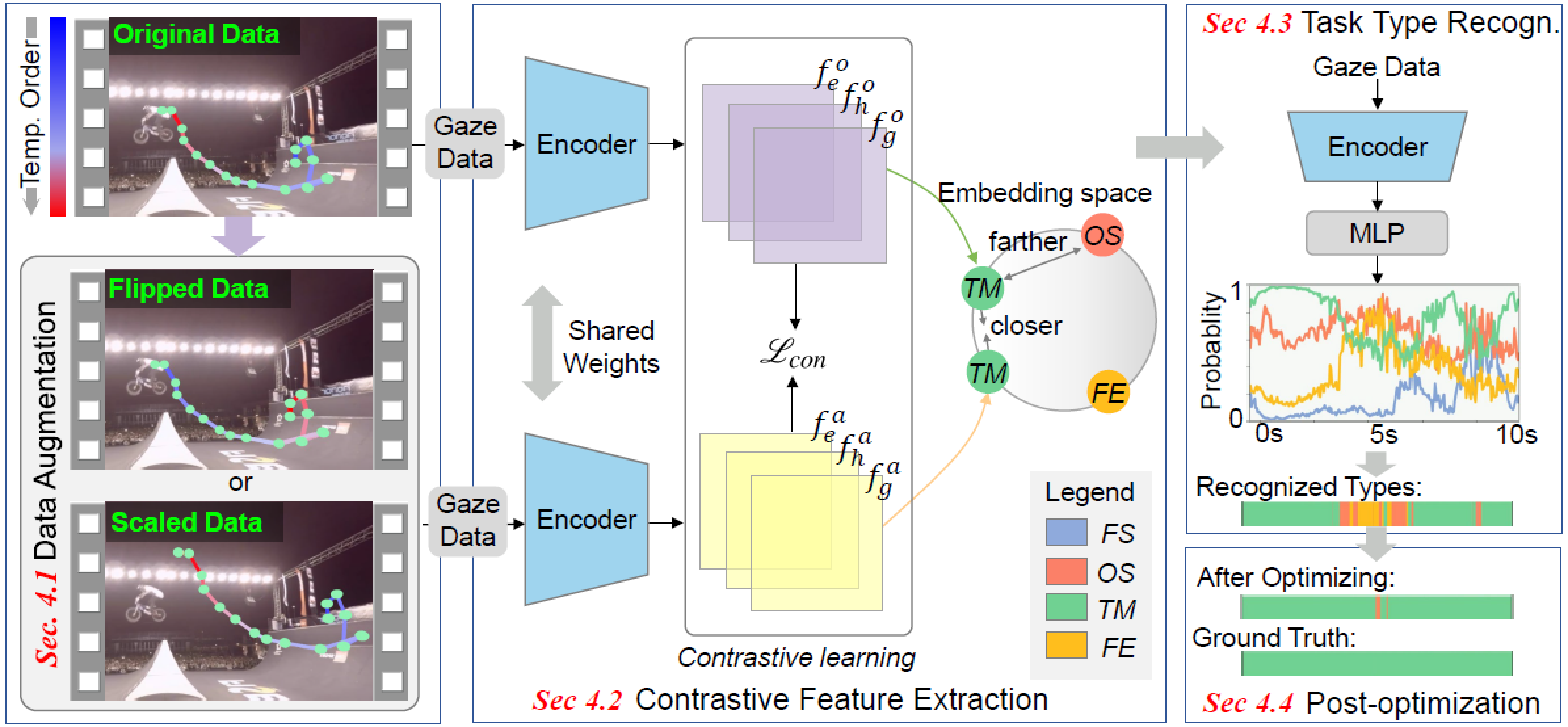
Task Type Name	FS	OS	TM	FE
Task Type Number	498	499	373	392
Task Type Duration (Total, min)	163.5	173.0	132.4	130.1
Duration Proportion (%)	27.3%	28.9%	22.1%	21.7%



(a) Task Type Number (b) Task Type Duration (c) Switch Count

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 is implemented to smooth the recognition results.



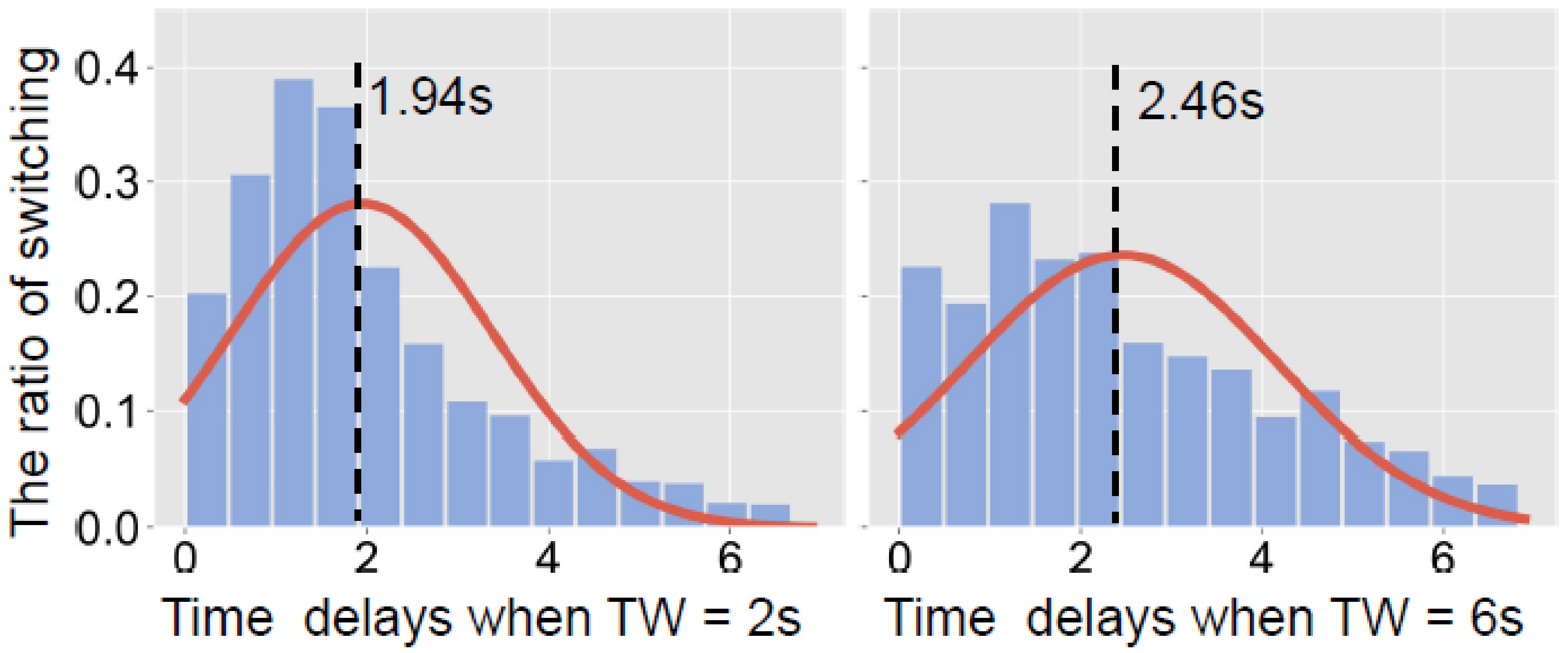
Overview of the proposed *TRCLP*

Experiment Results

Compare with state-of-the-art methods:

- Unlike other methods that perform best with a 6-second 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 the 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 0.5-second advantage.

	Time Window (s)	2s	6s	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%



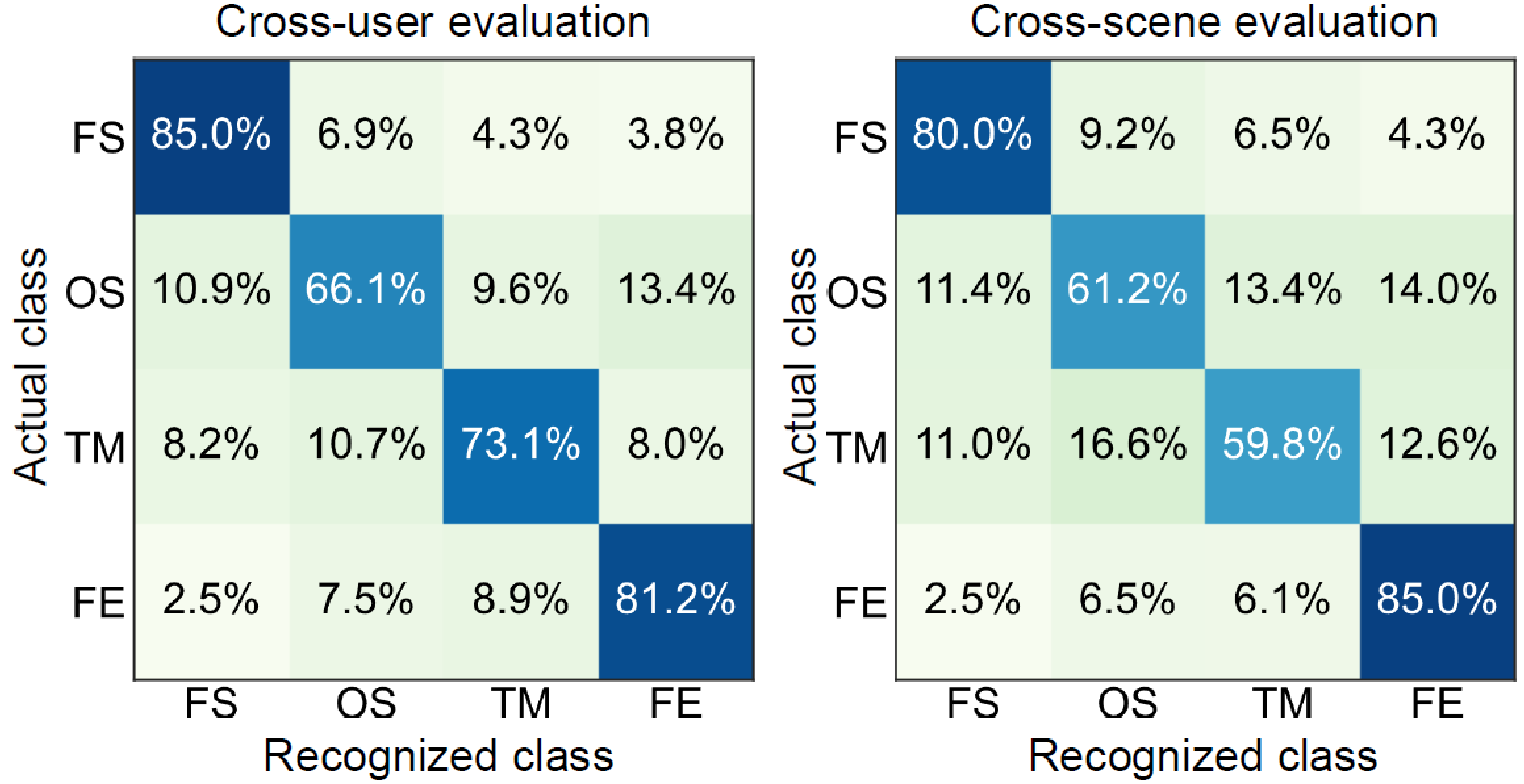
Experiment Results

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.

The confusion matrices of our method

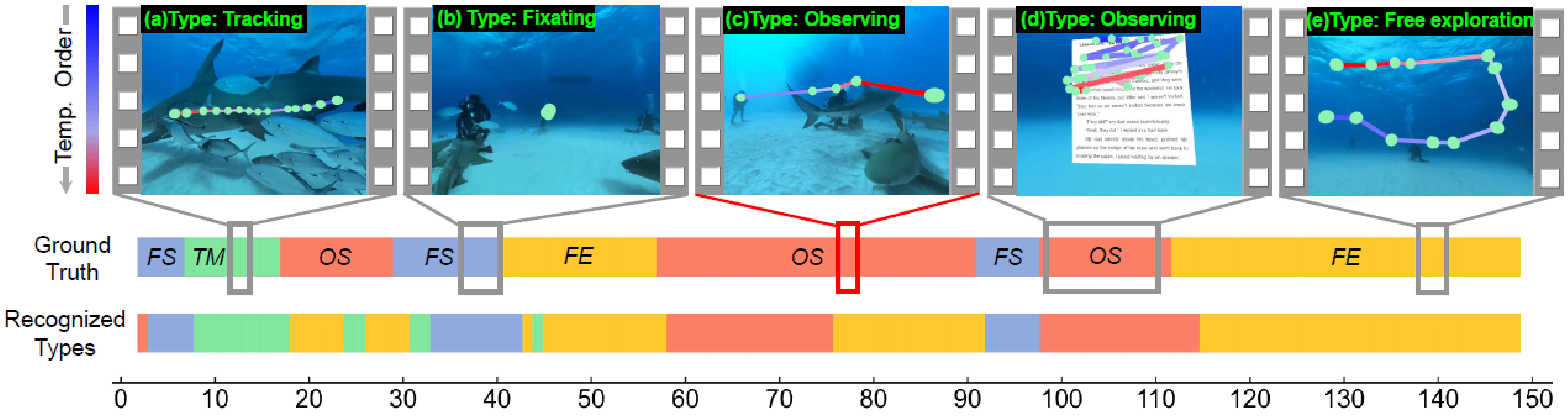


Ablation study of each component in our method

Baseline	DA	CL	POST	Cross-User	Cross-Scene
✓				71.8%	66.9%
✓	✓			72.1%	67.6%
✓	✓	✓		72.4%	67.9%
✓	✓	✓	✓	**76.1%**	**71.2%**

Experiment Results

This example shows the eye movement trajectories of a user while performing four task types in the VR video of the underwater world.



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

Thank you!

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