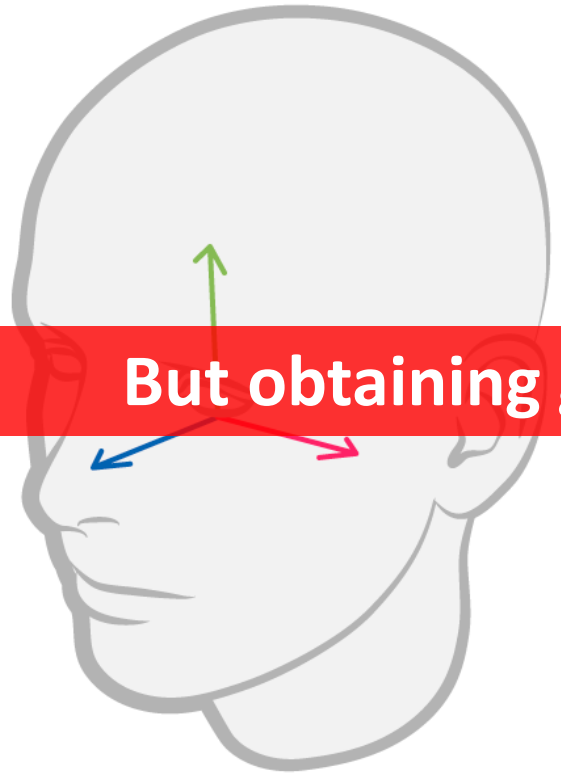
IEEE INFOCOM
IEEE International Conference on Computer Communications
10-13 May 2021 // Virtual Conference

# vGaze: Implicit Saliency-Aware Calibration for Continuous Gaze Tracking on Mobile Devices

**Songzhou Yang**, Yuan He, Meng Jin

School of Software and BNRist, Tsinghua University

# Eye movement and Gaze



**But obtaining gaze direction is not the final goal!**

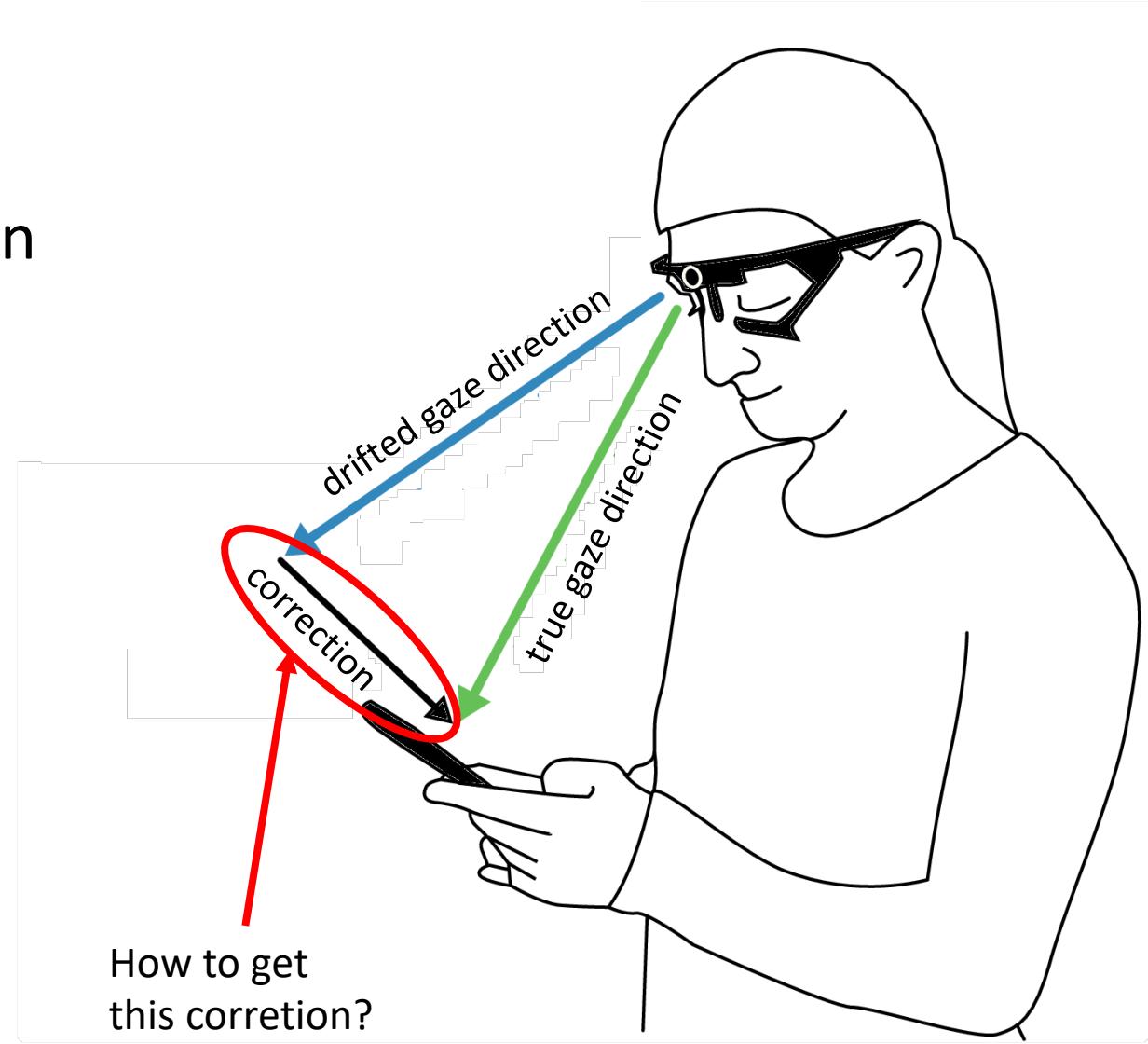
The gaze reflects where the user looks at.

**From the user's eye movement, we can infer the gaze direction.**

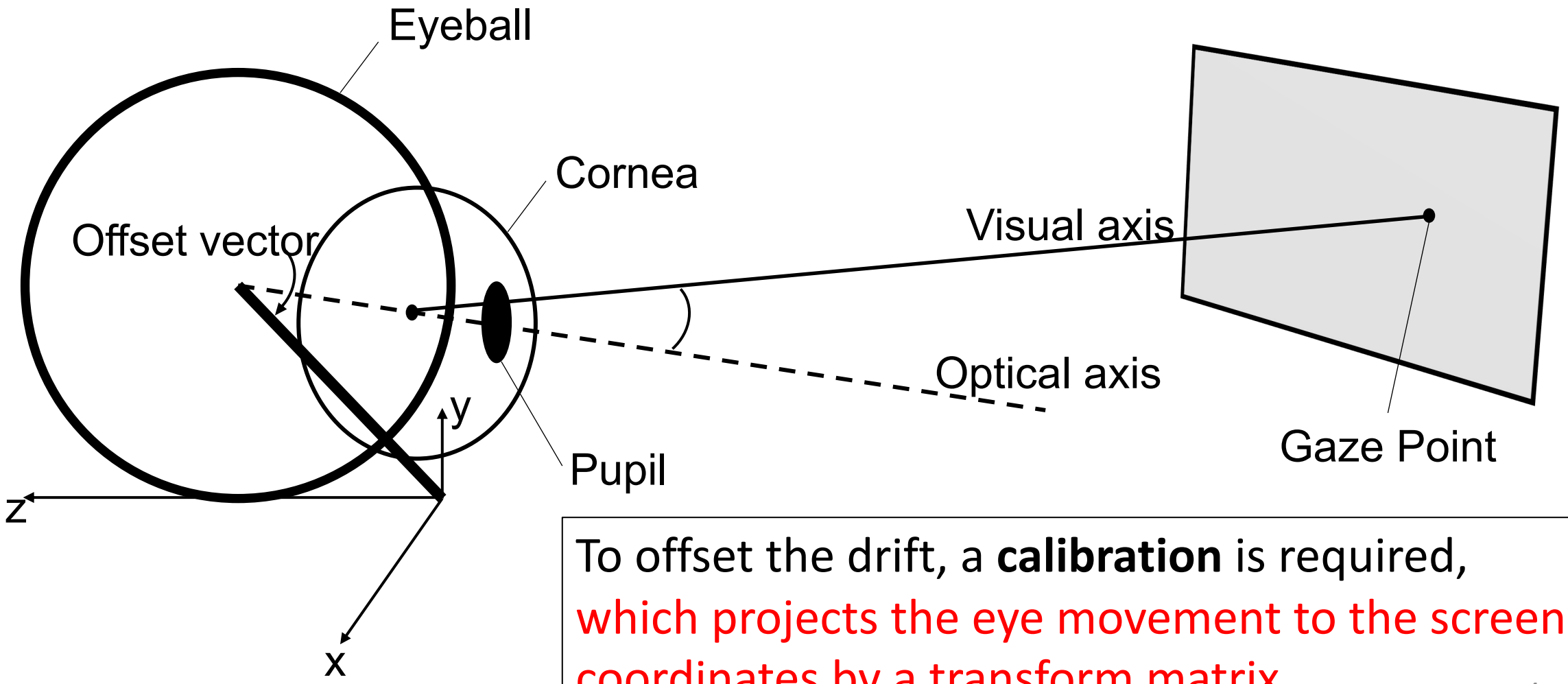
# Gaze Tracking Usage

- Gaze tracking usually acts as an interaction method.

- What we need is not the gaze direction, but the gaze on the screen!



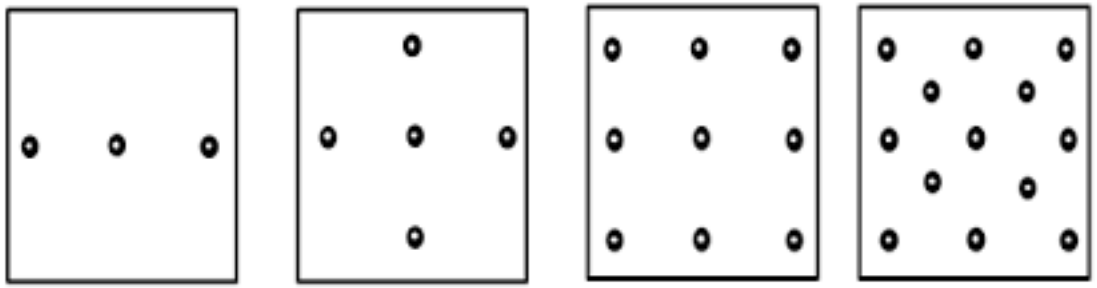
# Calibration



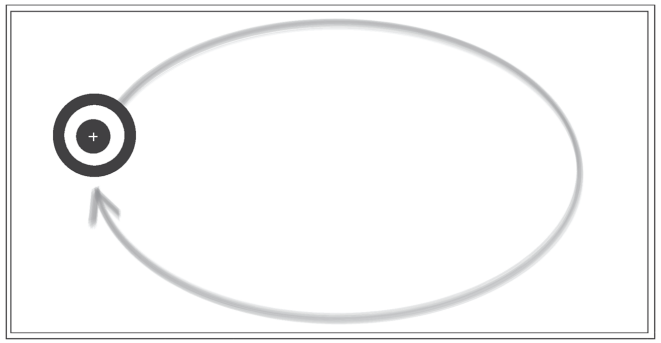

To offset the drift, a **calibration** is required, which projects the eye movement to the screen coordinates by a transform matrix.

# How the calibration is established?

- Traditional approaches require the user's cooperation to gaze at stimulus points at predefined coordinates on the screen, known as explicit calibration.



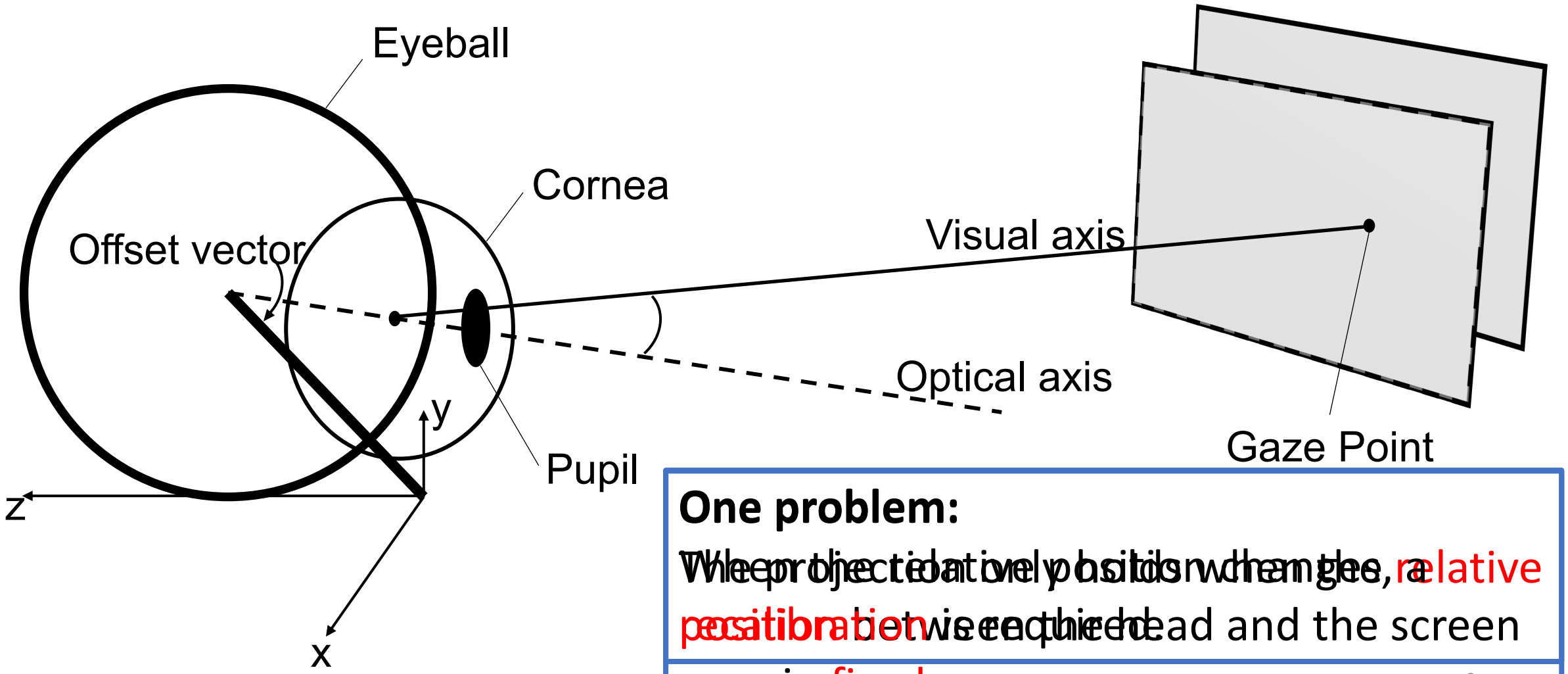
Dot-based Calibration



Pattern-based Calibration

**Effective, but …**

# Gaze Tracking




**One problem:**

~~The projection only holds when the~~ relative position between the head and the screen remain fixed.

# However…

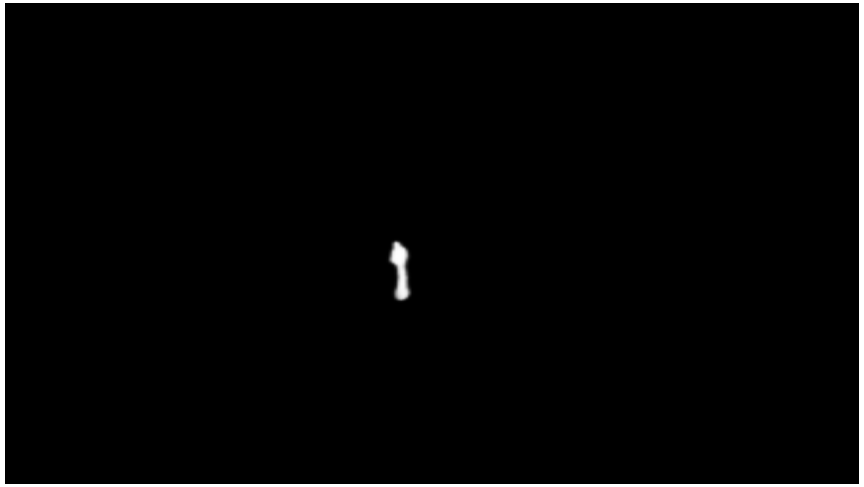
**Such a calibration process takes too much time.**
**Once the relative position changes, the re-calibration impairs the user's experiece heavily!**

# One Insight

Areas
draw attention

These areas are called saliency, which is a kind of visual information (i.e., distinctive color, intensity, orientation, objects, etc.)

# One Insight



**The Saliency Map**

In order to express saliency, the saliency map is generated.
Basically, the saliency map can be used as a kind of implicit stimuli for calibration.

# However…

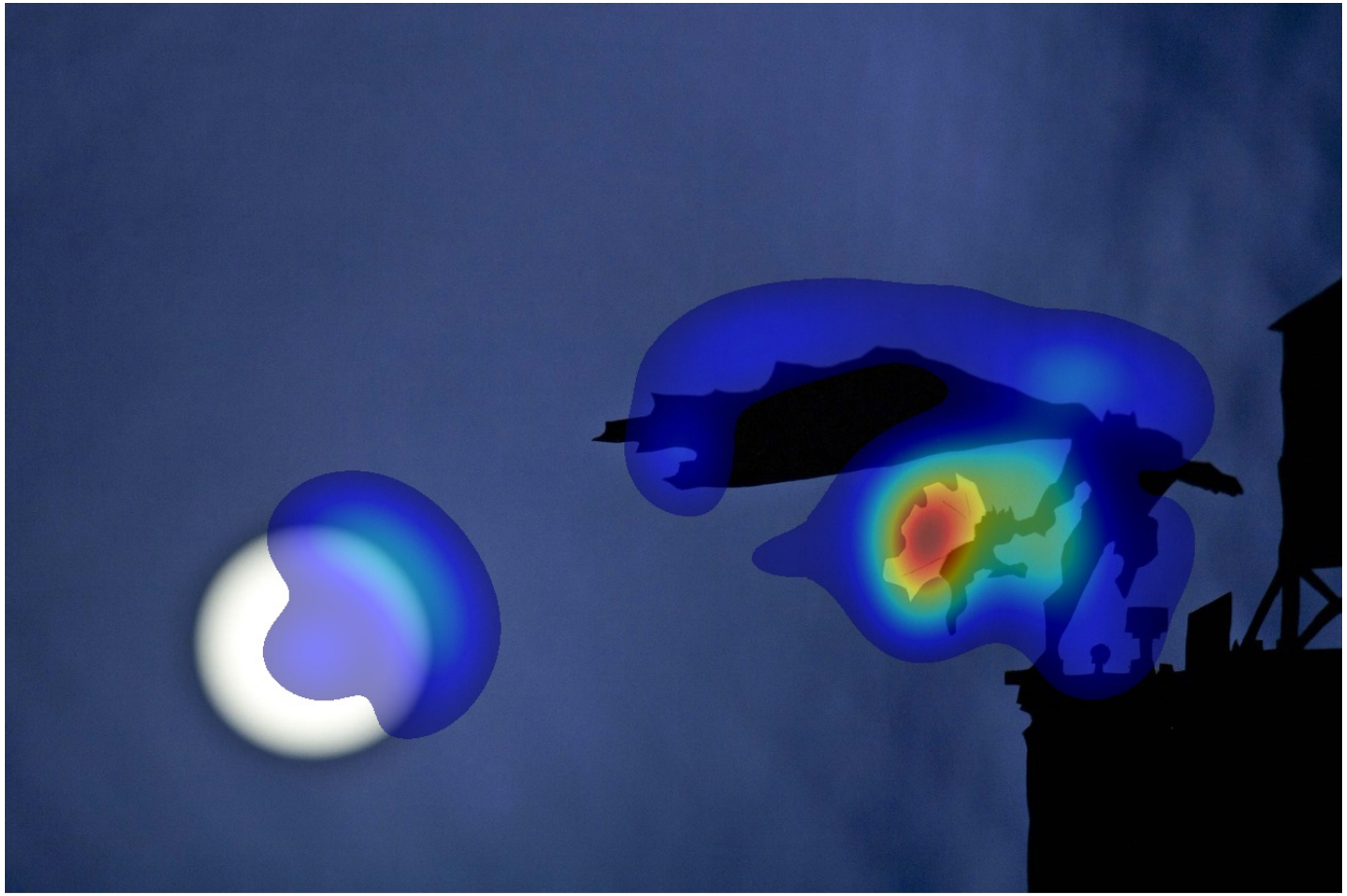
- Saliency is ambiguous in many frames.

For a complicated frames, there are always more than one areas act as saliency.

Moreover, these saliency play different r[illegible] in spatial and temporal d[illegible]ension.

# To better understand saliency

- The human attention mechanism.

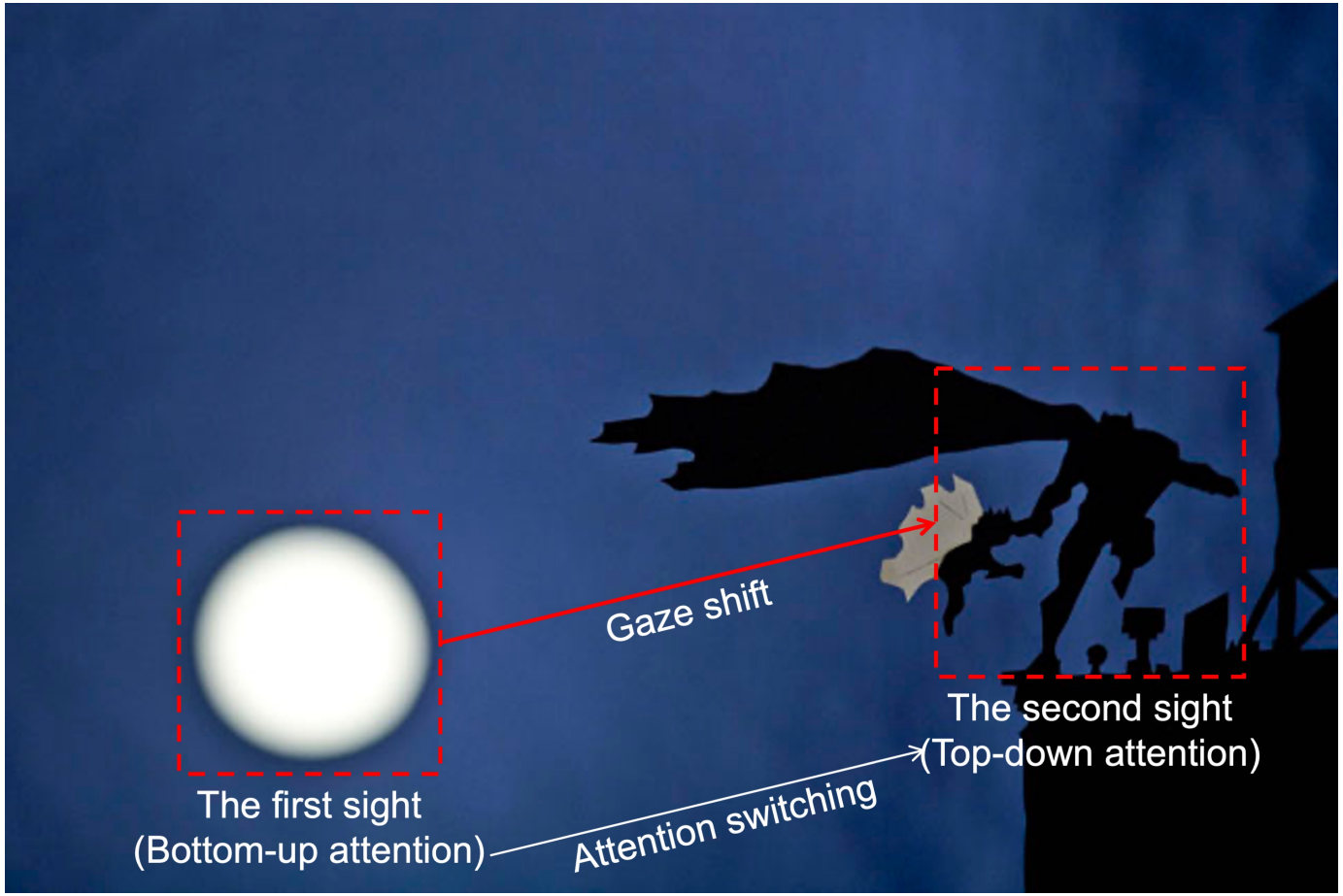


The user's attention is in the bottom-up mode during the first around 150ms after scene cuts.

Then, the user's attention enters the top-down mode, where the user's consciousness dominates the gaze.

# To better understand saliency

- Corresponding saliency.




## Bottom-up Saliency

salient because of their inherent properties relative to the background

e.g., luminosity, shape

## Top-down Saliency

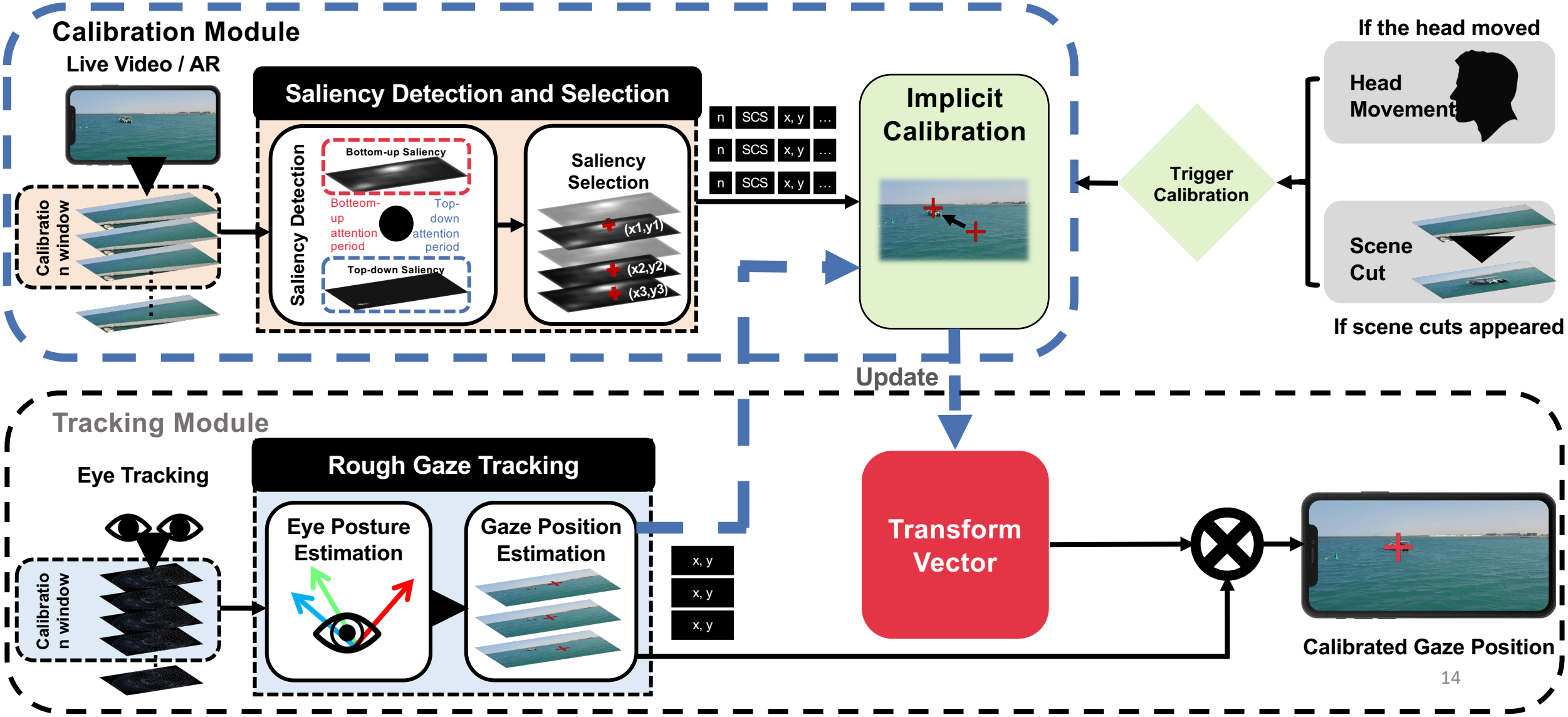
based on prior knowledge, willful plans, and current goals

Specifical kind of saliency should be used at specifical time.

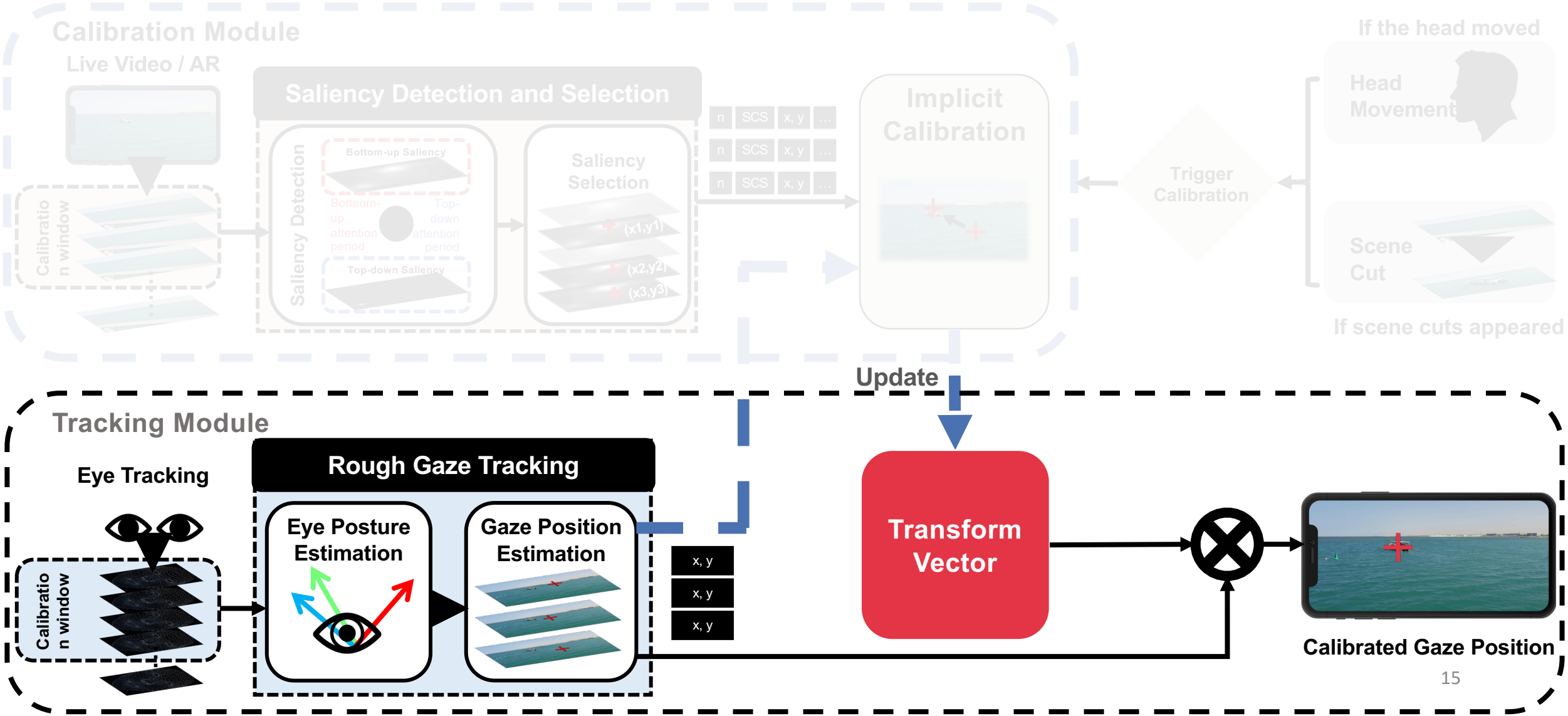
# Our Idea

**Leverage the temporally and spatially dependent relation between the saliency and the user's attention.**

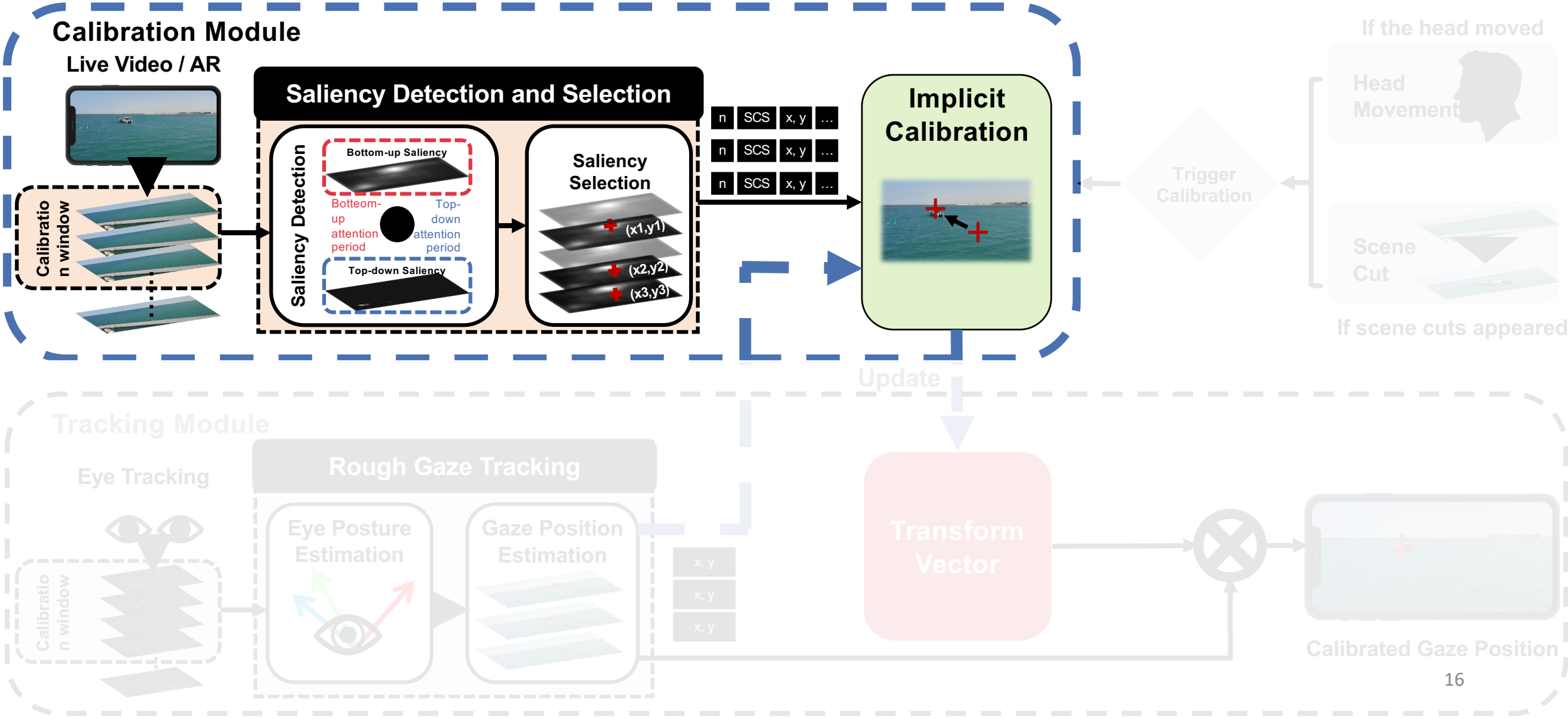
# Design


Calibration Module
Live Video / AR
Calibratio n window
Saliency Detection and Selection
Saliency Detection
Bottom-up Saliency
Botteom-up attention period
Top-down attention period
Top-down Saliency
Saliency Selection
(x1,y1)
(x2,y2)
(x3,y3)
n SCS x, y …
n SCS x, y …
n SCS x, y …
Implicit Calibration
Trigger Calibration
If the head moved
Head Movement
Scene Cut
If scene cuts appeared
Update
Tracking Module
Eye Tracking
Calibratio n window
Rough Gaze Tracking
Eye Posture Estimation
Gaze Position Estimation
x, y
x, y
x, y
Transform Vector
Calibrated Gaze Position

# Design

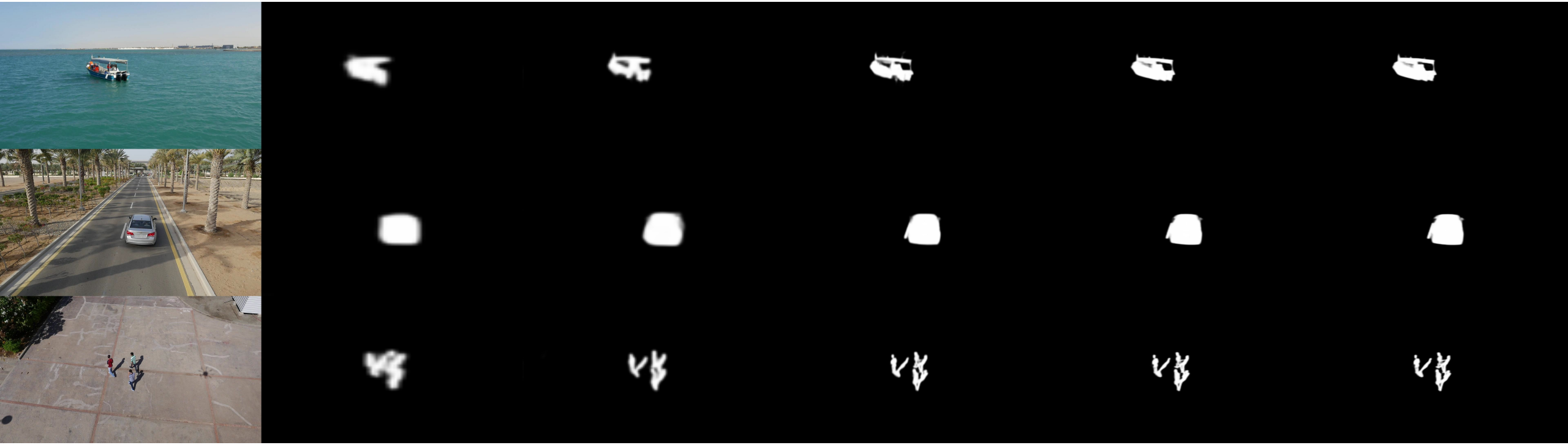

Calibration Module
Live Video / AR
Calibratio n window
Saliency Detection and Selection
Saliency Detection
Bottom-up Saliency
Botteom-up attention period
Top-down attention period
Top-down Saliency
Saliency Selection
(x1,y1)
(x2,y2)
(x3,y3)
n SCS x, y …
n SCS x, y …
n SCS x, y …
Implicit Calibration
Trigger Calibration
If the head moved
Head Movement
Scene Cut
If scene cuts appeared
Update
Tracking Module
Eye Tracking
Calibratio n window
Rough Gaze Tracking
Eye Posture Estimation
Gaze Position Estimation
x, y
x, y
x, y
Transform Vector
Calibrated Gaze Position

# Design


Calibration Module
Live Video / AR
Calibratio n window
Saliency Detection and Selection
Saliency Detection
Bottom-up Saliency
Botteom-up attention period
Top-down attention period
Top-down Saliency
Saliency Selection
(x1,y1)
(x2,y2)
(x3,y3)
n SCS x, y …
n SCS x, y …
n SCS x, y …
Implicit Calibration
Trigger Calibration
If the head moved
Head Movement
Scene Cut
If scene cuts appeared
Update
Tracking Module
Eye Tracking
Calibratio n window
Rough Gaze Tracking
Eye Posture Estimation
Gaze Position Estimation
x, y
x, y
x, y
Transform Vector
Calibrated Gaze Position

# Design

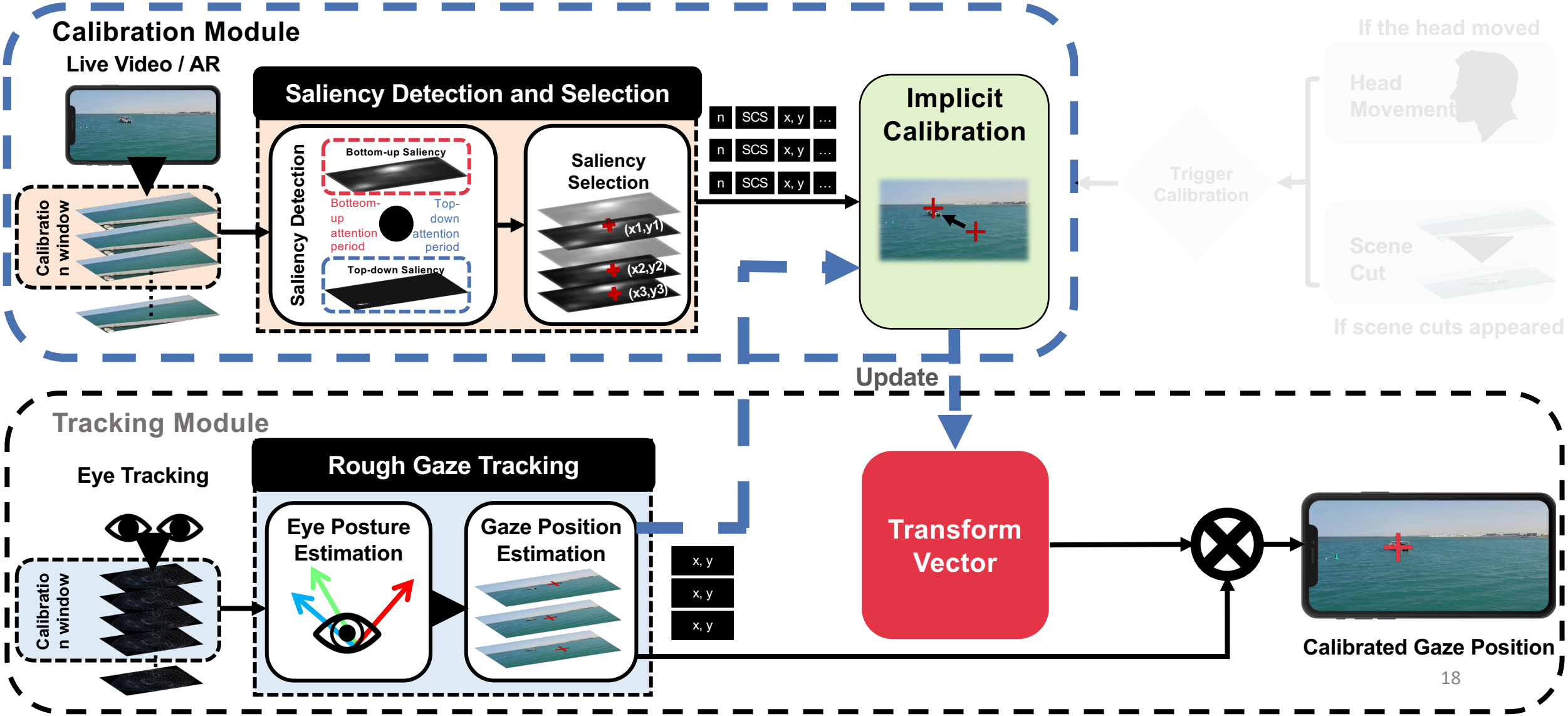
- Original frame and saliency map with different resolution



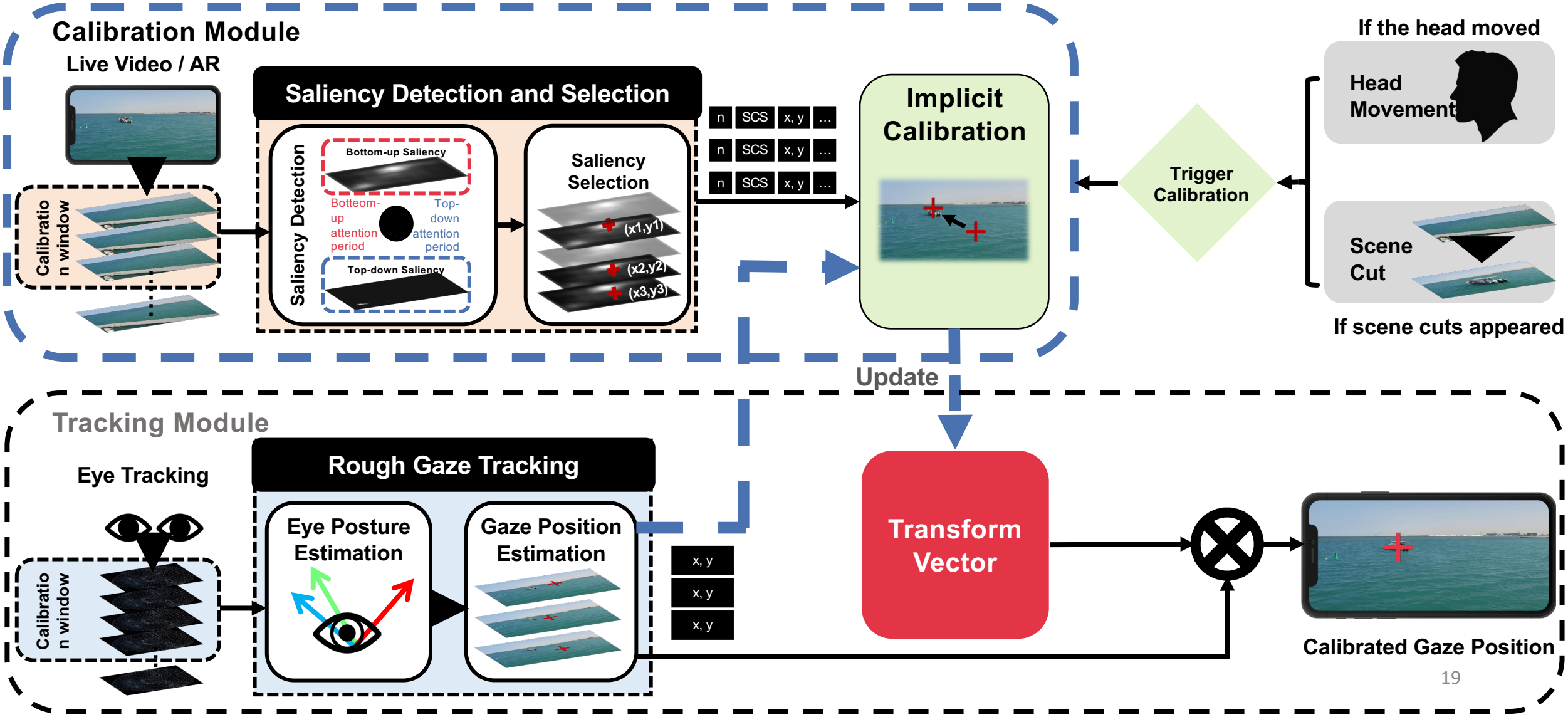
68x68 160x90 320x180 640x360 1280x720

**Resulotion does not influence the detection of saliency**

# Design

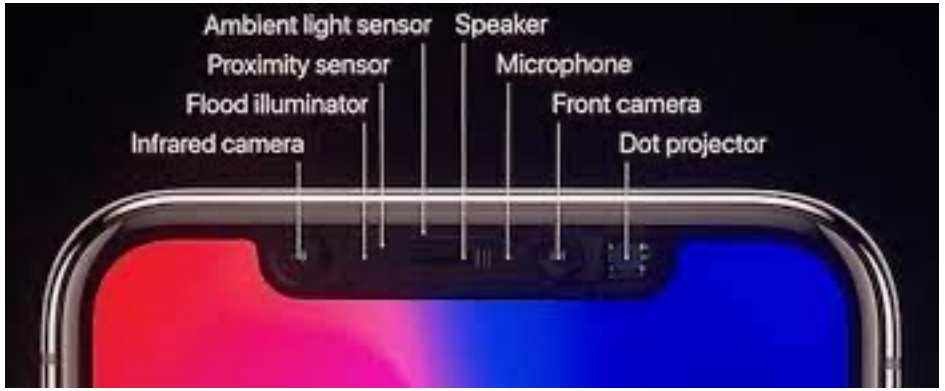

Calibration Module
Live Video / AR
Calibratio n window
Saliency Detection and Selection
Saliency Detection
Bottom-up Saliency
Botteom-up attention period
Top-down attention period
Top-down Saliency
Saliency Selection
(x1,y1)
(x2,y2)
(x3,y3)
n SCS x, y …
n SCS x, y …
n SCS x, y …
Implicit Calibration
If the head moved
Head Movement
Trigger Calibration
Scene Cut
If scene cuts appeared
Update
Tracking Module
Eye Tracking
Calibratio n window
Rough Gaze Tracking
Eye Posture Estimation
Gaze Position Estimation
x, y
x, y
x, y
Transform Vector
Calibrated Gaze Position

# Design


Calibration Module
Live Video / AR
Calibratio n window
Saliency Detection and Selection
Saliency Detection
Bottom-up Saliency
Botteom-up attention period
Top-down attention period
Top-down Saliency
Saliency Selection
(x1,y1)
(x2,y2)
(x3,y3)
n SCS x, y …
n SCS x, y …
n SCS x, y …
Implicit Calibration
Trigger Calibration
If the head moved
Head Movement
Scene Cut
If scene cuts appeared
Update
Tracking Module
Eye Tracking
Calibratio n window
Rough Gaze Tracking
Eye Posture Estimation
Gaze Position Estimation
x, y
x, y
x, y
Transform Vector
Calibrated Gaze Position

# Implementation

- iPhone XS Max
  - ARKit & TrueDepth Camera
    - Eye movement tracking
  - IMU Sensors
    - Phone posture detection





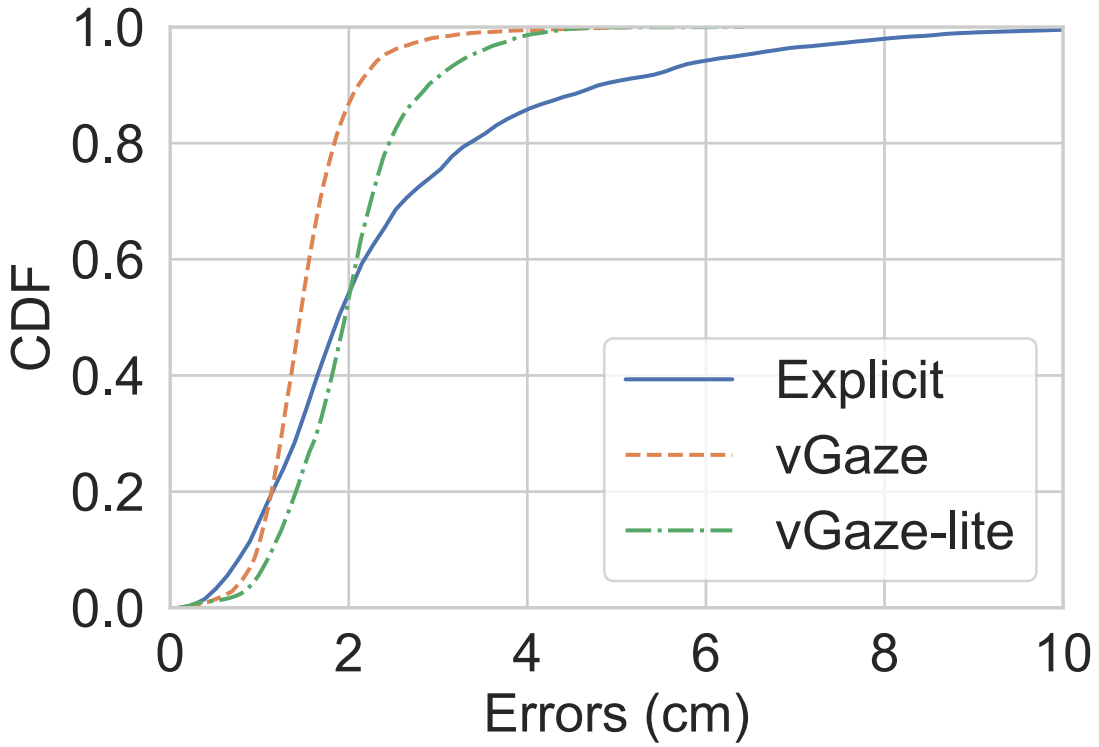
# Evaluation

- Evaluation setups:
  - 6 videos (from EyeTrackUAV dataset)
  - 10 volunteers
    - 5 males and 5 females
    - ages vary from 8 to 72 years old.

| Title | Duration | Resolution | Sample Rate | Total frames |
|---|---|---|---|---|
| bike3 | 14s | 1280*720 | 30fps | 432 |
| boat6 | 27s | 1280*720 | 30fps | 804 |
| boat8 | 23s | 1280*720 | 30fps | 684 |
| building5 | 16s | 1280*720 | 30fps | 480 |
| car6 | 73s | 1280*720 | 30fps | 2194 |

# Evaluation

- Overall tracking errors: 1.51cm

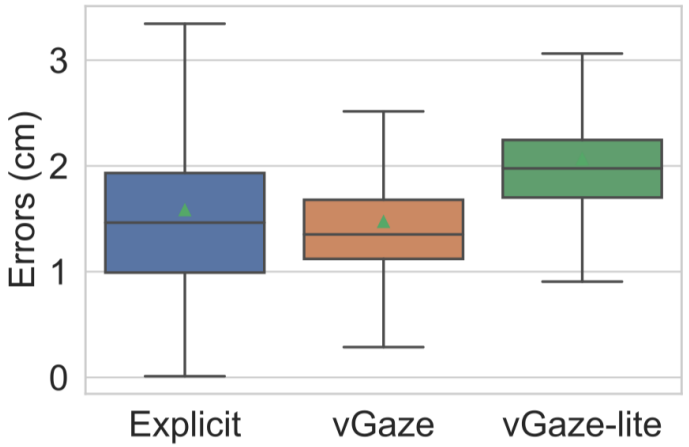



Explict: explicit calibration with five dots
vGaze: our solution
vGaze-lite: only bottom-up saliency is used

Three scenario are involved, static (where the user stay static), dynamic (where the user is asked to move), natural (where the user's movement is not constrained)

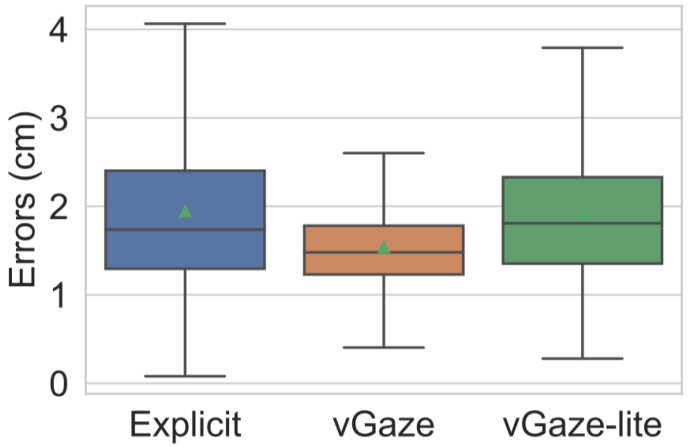
# Evaluation

- Errors on three different scenarios



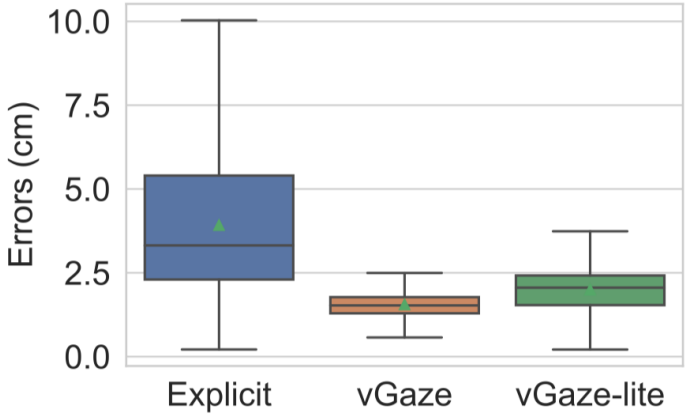

(a) Static

vGaze is comparable with explict calibration on errors in static scenarios without interruption.




(b) Natural

vGaze is better than explicit calibration in other two senarios.

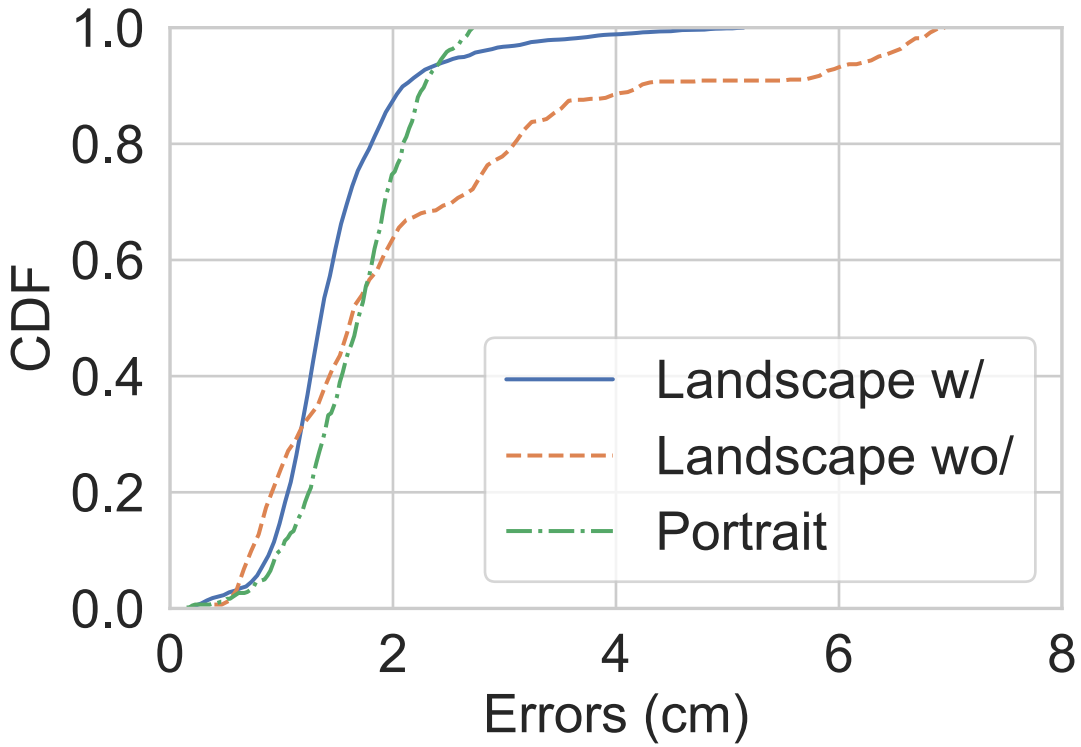



(c) Dynamic

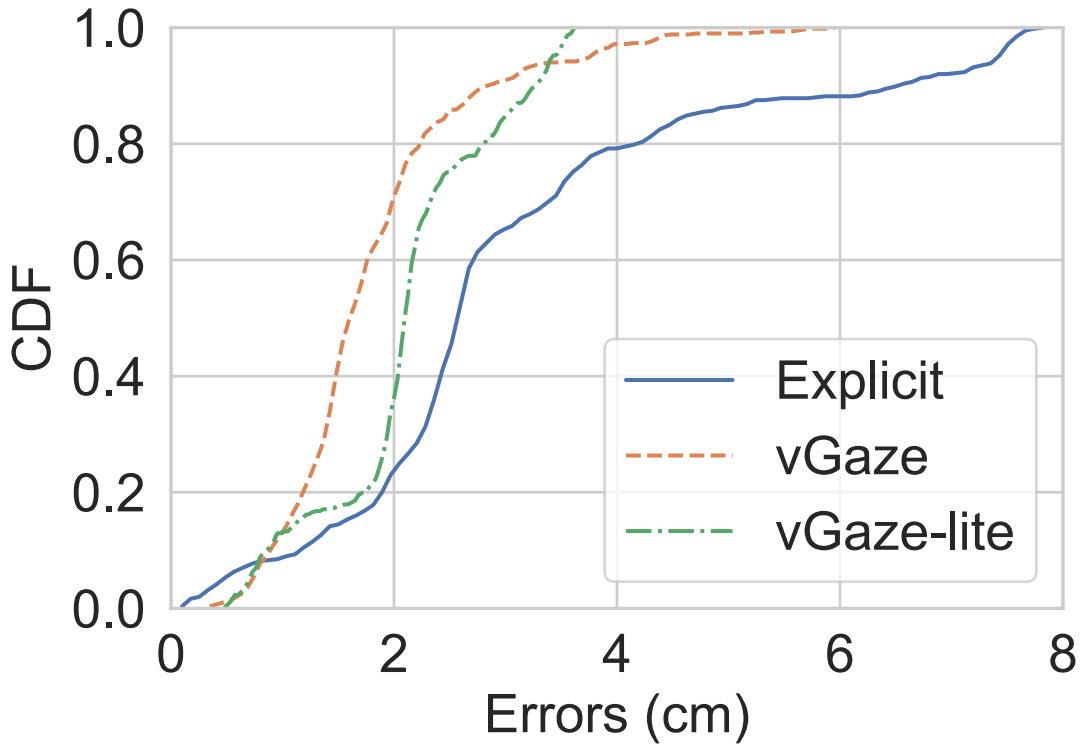
vGaze beats vGaze-lite in all senarios

# Evaluation

- Landscape v.s. Portrait

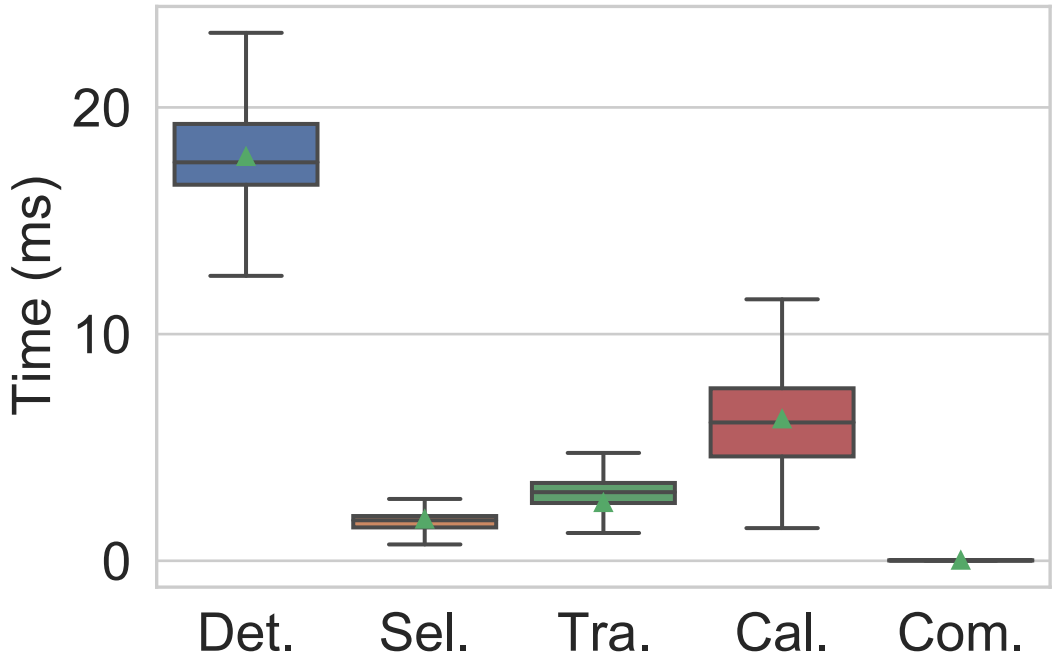



- Phone held in hand



# Evaluation

- Time elapsed by different modules




- Saliency Detection
- Saliency Selection
- Rough Gaze Tracking
- Calibration
- Compensation of rough gaze tracking and transform vector

The average total time consumed by saliency detection and selection is 19.66 ms for a frame, which is much shorter than the frame display interval 33.33ms of 30 FPS video/AR.

# In Summary

- With the insight of the **temporal** and **spatial** relation between the gaze and the visual saliency, we present the design and implementation of vGaze, implicit saliency-aware calibration for **continuous gaze tracking on mobile devices**.
  - Bottom-up saliency & Top-down saliency
  - High accuracy & Low latency

My email: yangsz18@mails.tsinghua.edu.cn

# Thanks For listening

# Q & A