Visionary Co-Driver: LLMs Enhance Driver Perception of Potential Risks with AR-HUD

Enable Blind LLMs with Vision

---My Summer Research Summary

王杰，04/06/2024

Advisor: 向为教授 Prof. Wei Xiang
Mentor: 黄莹莹 Yingying Huang

Github: Summer_Research_2023

Visionary Co-Driver: LLMs Enhance Driver Perception of Potential Risks with AR-HUD

Content

1. Project Intuition
2. Model Selection
3. System Design
4. User Study
5. Failure and Success
6. Research Reflection

Intuition:
Corner cases are countless, but LLM has logic

Road scene understanding is a relative complex and important topic in autonomous driving field.

Traditional approach like **driving risk field** or **object detection** is useful and practical, but they can’ t cover all the possibilities in the real-world scenario (Too much annoying if-statements).

Solving it in a brand-new way is my intuitive study interest, because I saw the potential capability within LLM to be a co-pilot, helping drivers to recognize and avoid those potential risk.

HOWEVER, THERE IS A HUGE GAP BETWEEN LOGIC AND REALITY: *MODALITY DIFFERENCE*

The ChatGPT can 'think 'through text conversation, but it doesn’ t have internal capability of perception.

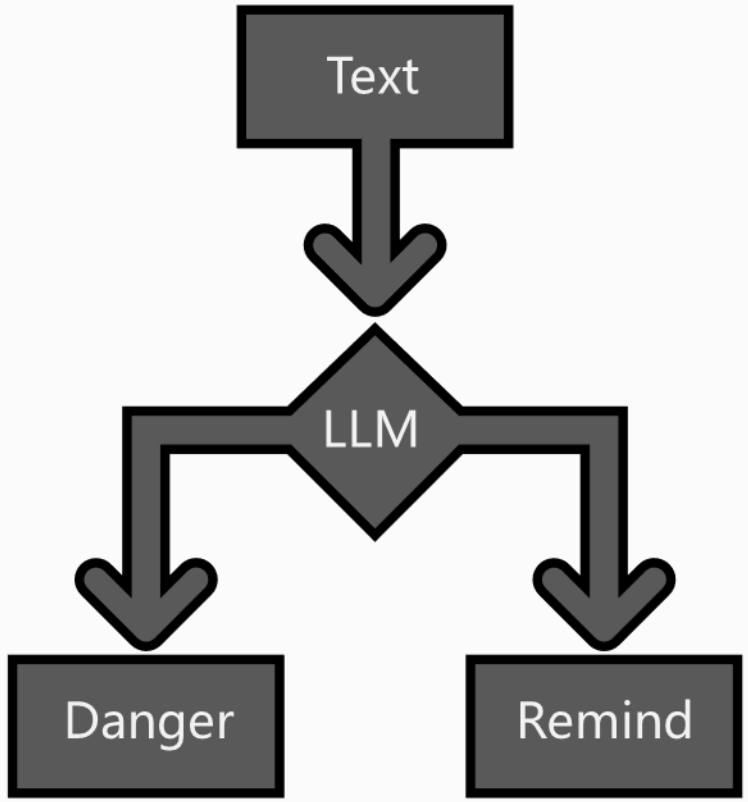
Therefore, we need to provide a tool for it to "read" the road (**Road Scene Understanding).**

“有啥子在我前头啊？看不着啊”

Visionary Co-Driver

OUR SOLUTION: COMBINE SOTA MODELS TOGETHER, OBTAIN ON-ROAD INFORMATION

A vision2text cascaded connected model, helping the LLM to detect the road



What is vision?

Video + Audio

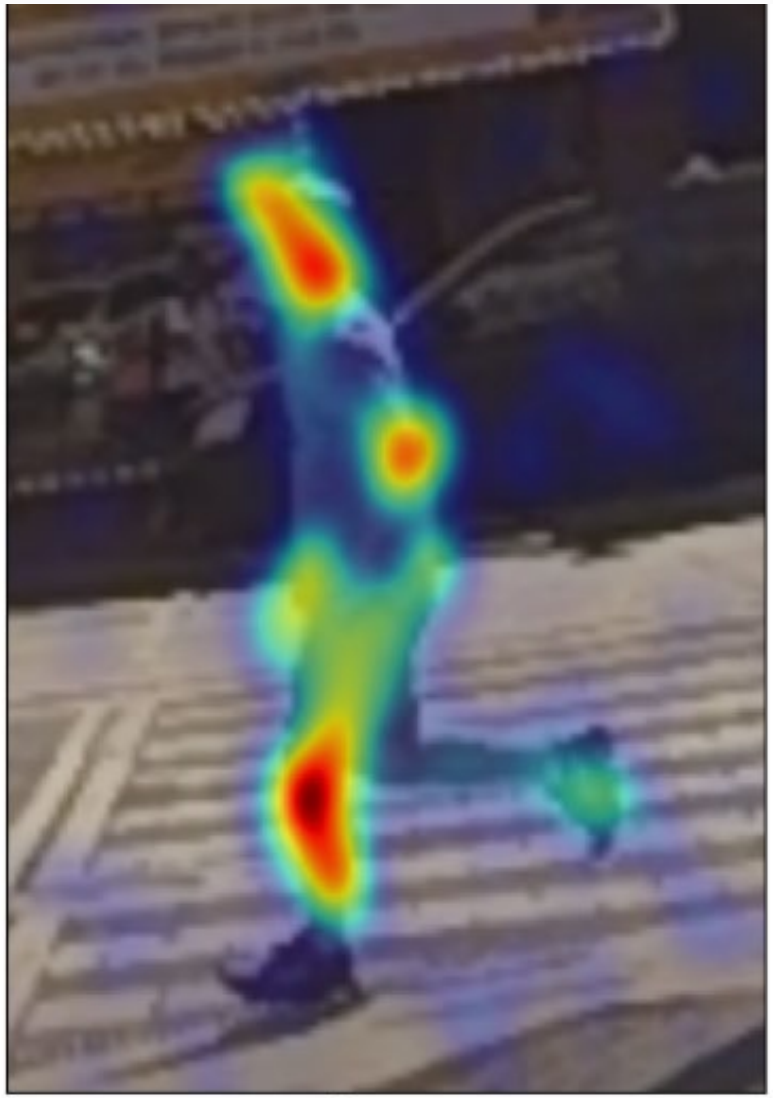
- Because we focus on driving in-car scenario, we skip audio part

Video == a sequence of frame

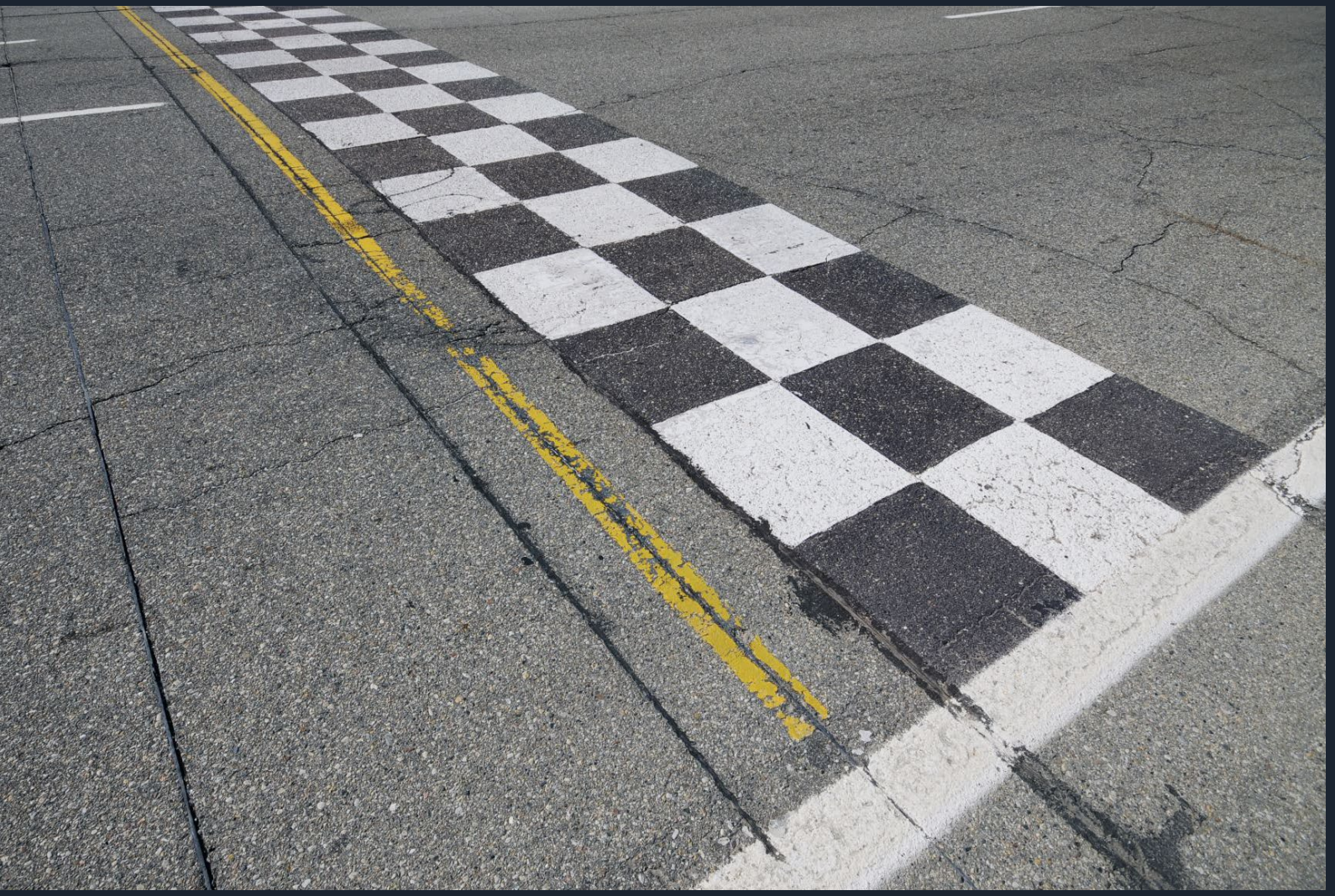
So, we need to firstly know what happens within a frame

WHAT IS THE VISION FEATURE OF A PEDESTRIAN?

- Location: road or sidewalk? Left or right side?
- Distance: near or far from us?
- Speed: fast or slow?
- Type: old or young?



Visionary Co-Driver



HOW TO OBTAIN THE ROAD / SIDEWALK INFORMATION?

OUR ANSWER: SAM (NOT UNCLE)

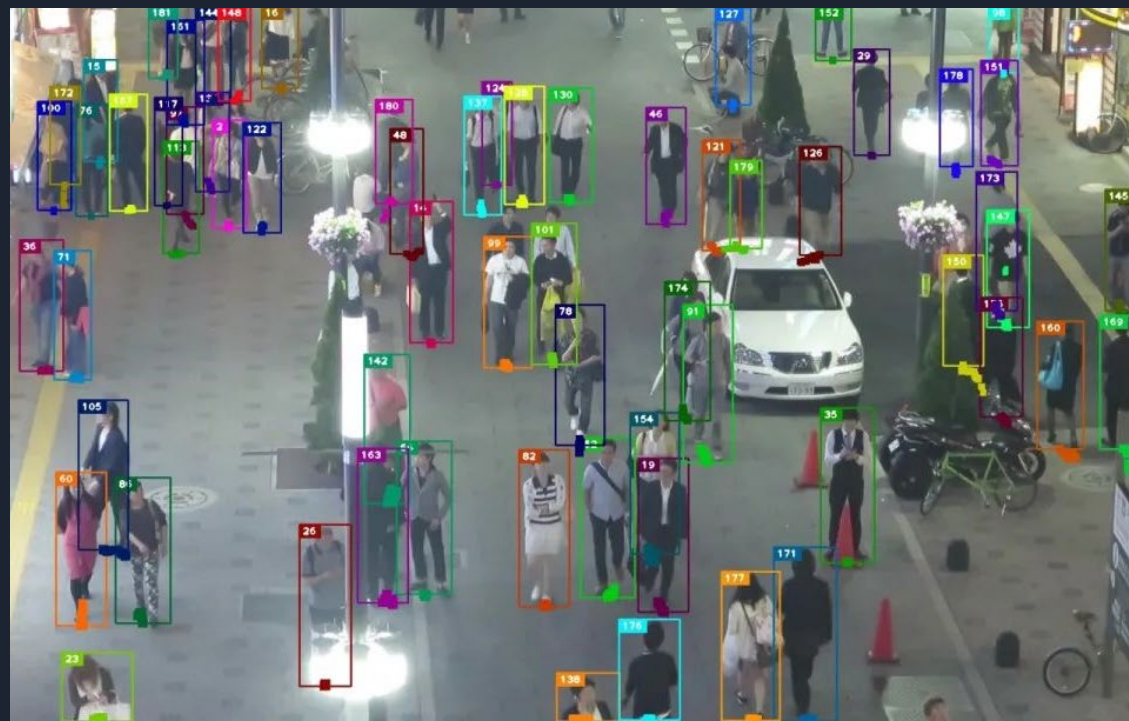
Visionary Co-Driver

Multiple Object Tracking Yolov8+ByteTrack

[2023 SOTA]

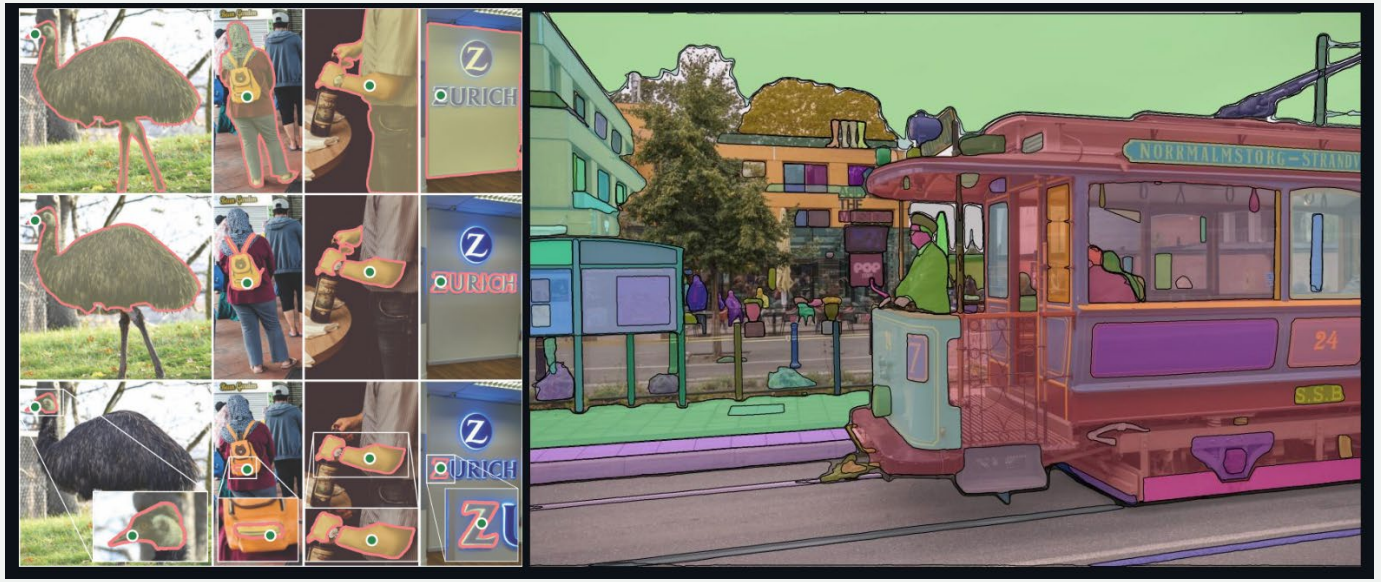
Official Support of the classic and powerful real-time CV:
You Only Look Once

Fastest with GPU acceleration



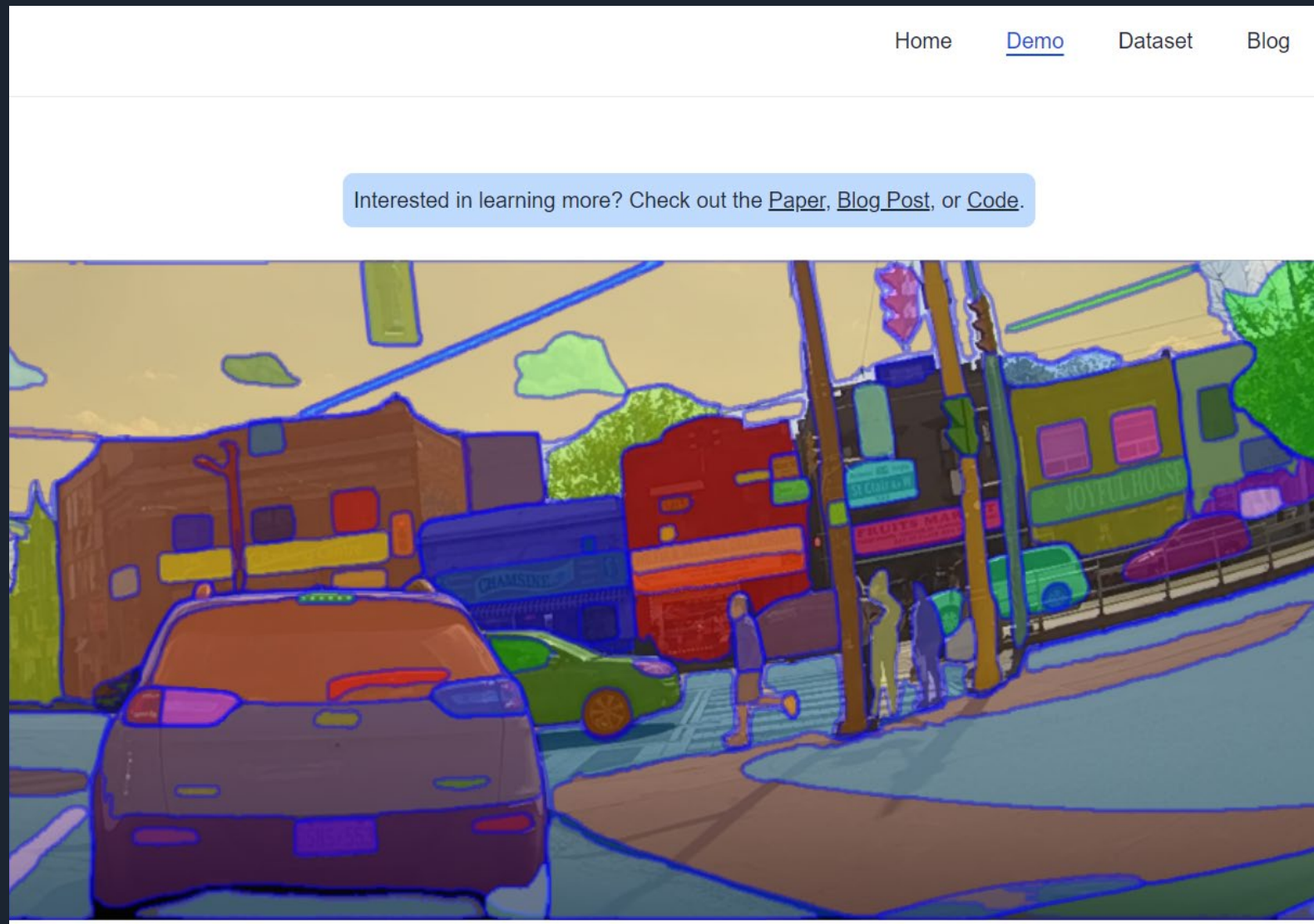
Segmentation: Segment Anything(SAM)

- [2023SOTA]
- Transformer based unlabeled segmentation, SOTA zero-shot segmentation



Divide the sidewalk from road

After segment everything, it’ s clear that all the sidewalks are shown in the image.



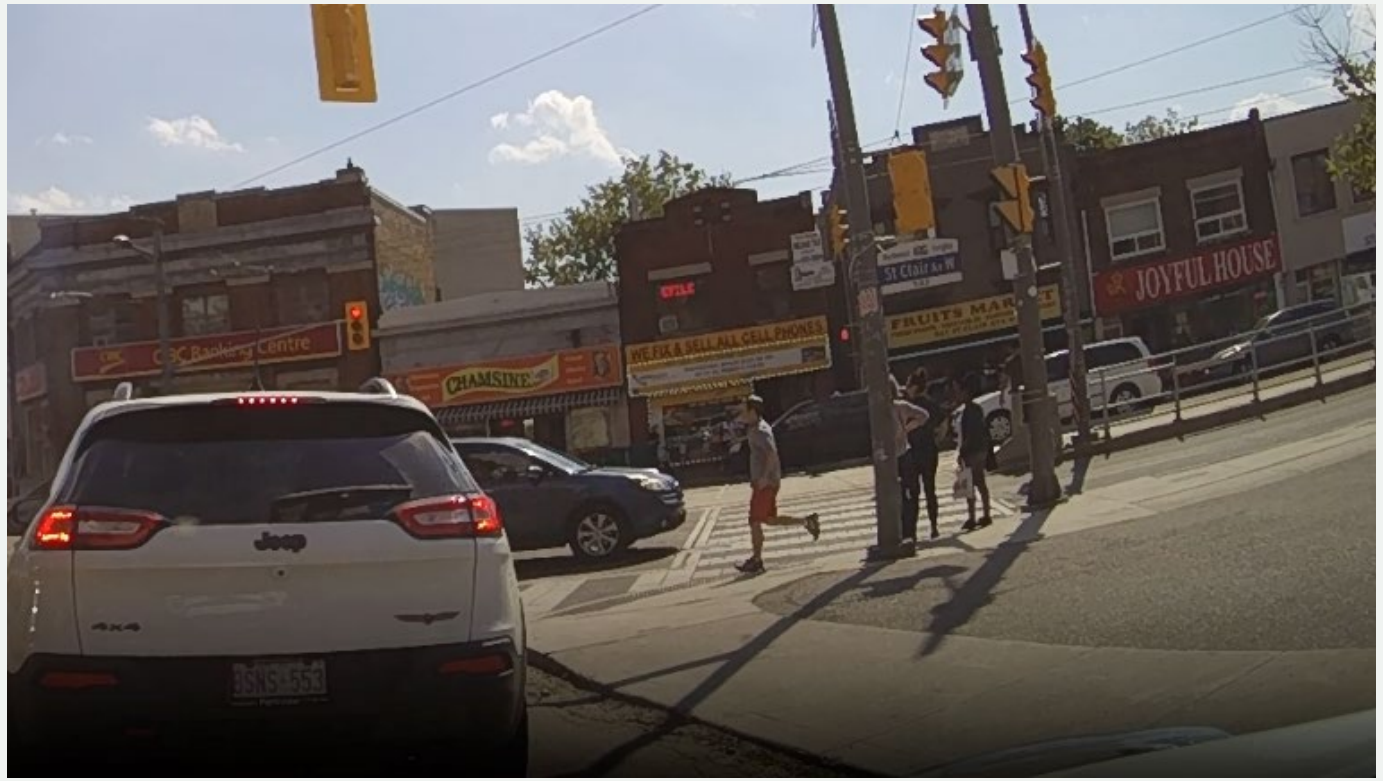
But segment everything is expensive!

How can we obtain the mask information easily from SAM?

Take a review on SAM:

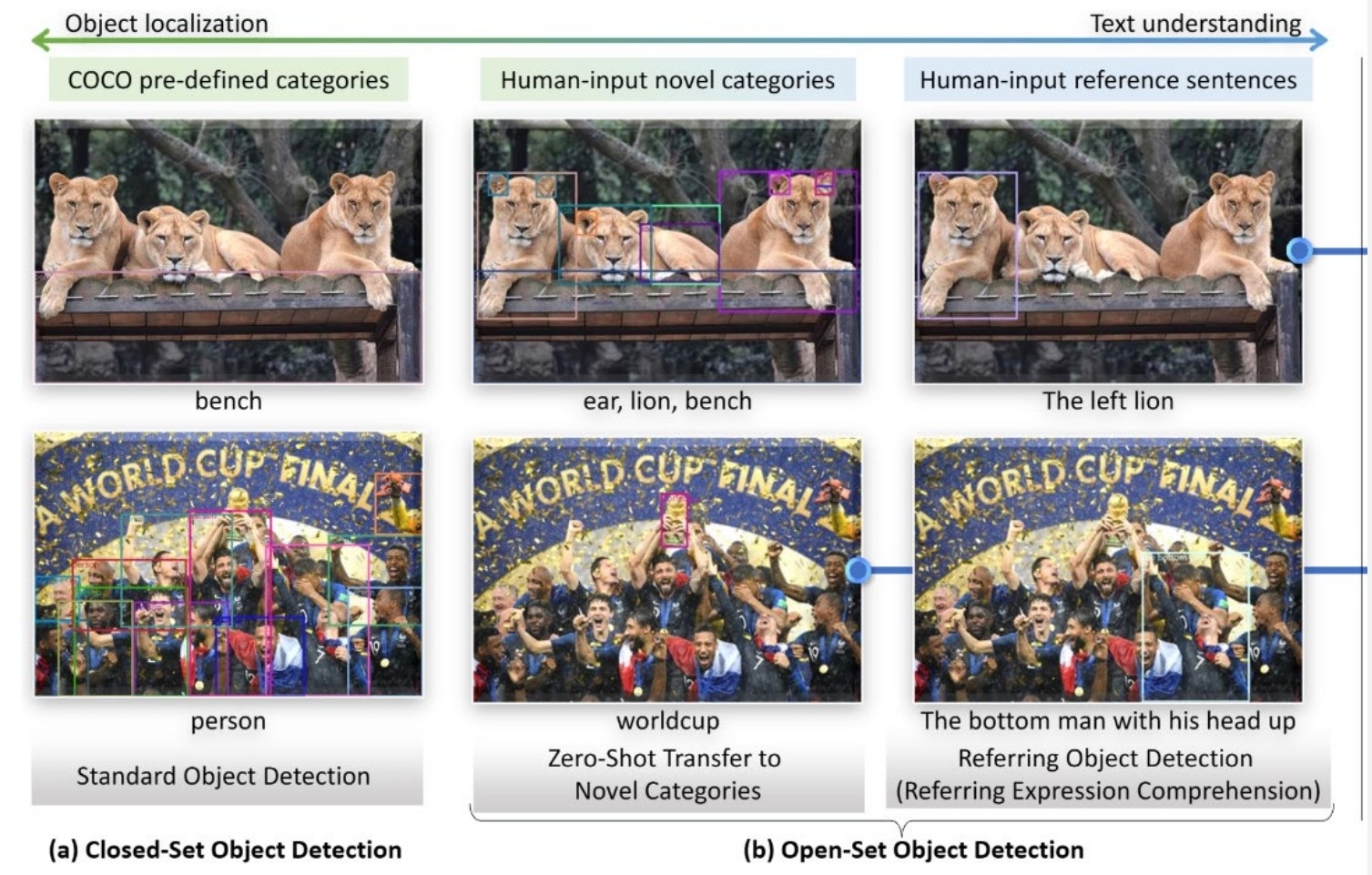
- can generate pixel level binary mask of the image

- Can be prompted by both point and box, including positive and negative prompt



Object detection: Grounding DINO

- [2022SOTA]
- Transformer based knowledge distillation, SOTA zero-shot detection.
- can generate bbox of the detected object using monocular image



Object detection: Grounding DINO

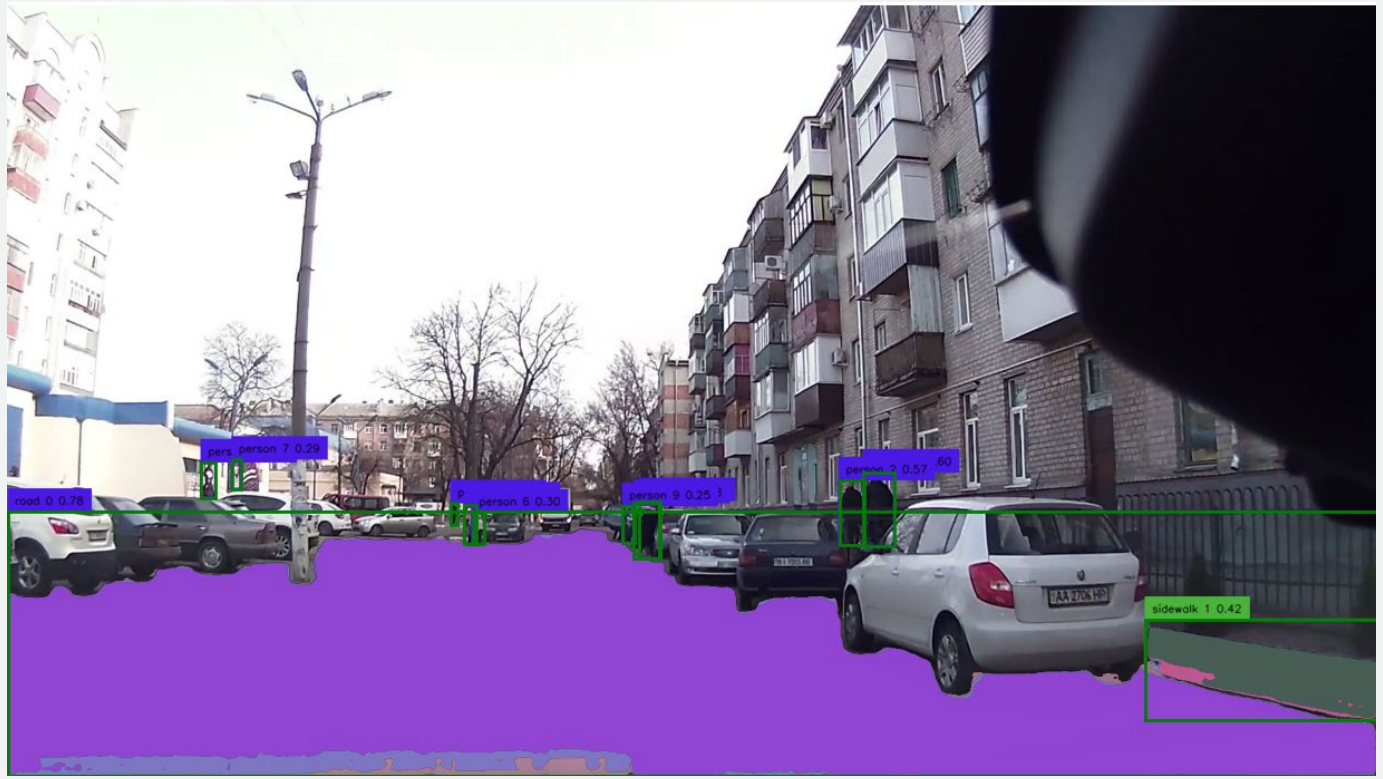
- Can be prompted based on nature language

- The gDINO can successfully recoginze the road & sidewalk using contrast prompting

- Pedestrian detection is separate, optimizing the accuracy

Take a review on SAM:

Combining with Grounding DINO, then we realize a SOTA semantic segmentation model



Now, how can we know the distance from pedestrian to us?

In industry practice:
Radar & Lidar Sensor!

But: we are poor, and we only have monocular video…

Visionary Co-Driver

ANSWER: MONOCULAR DEPTH PREDICTION MODAL!

Monocular Computer Vision is powerful these days!

Depth estimation: Dense Prediction Transformer(DPT)

- [2021SOTA]
- Transformer based monocular depth estimation
- can generate image-size NumPy array indicating the depth value of each pixel
- Based on it, we obtained the estimated distance from dashcam to object

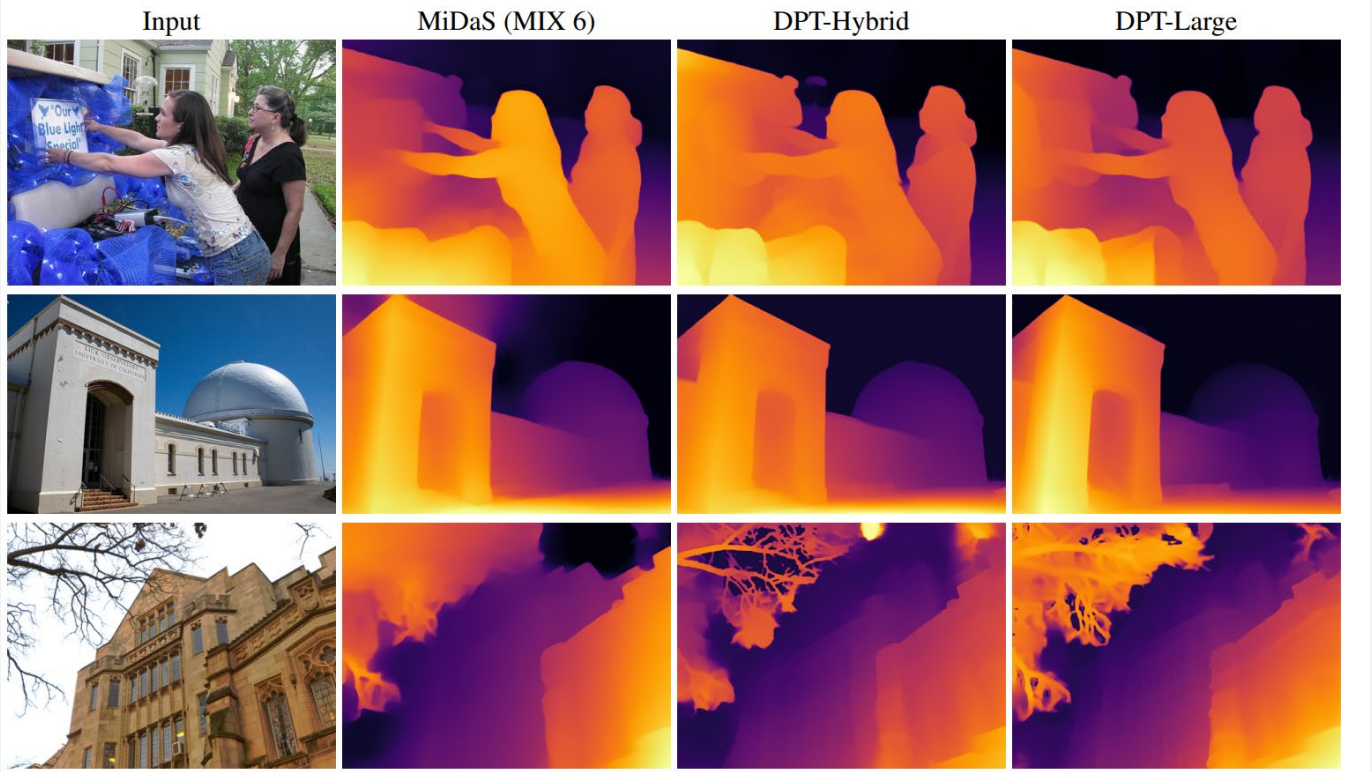
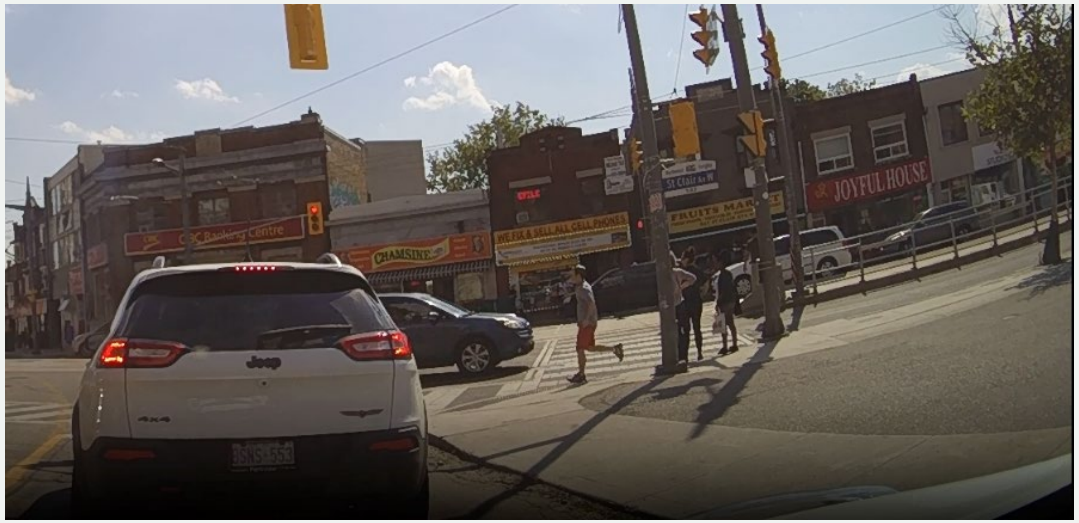


Figure 2. Sample results for monocular depth estimation. Compared to the fully-convolutional network used by MiDaS, DPT shows better global coherence (e.g., sky, second row) and finer-grained details (e.g., tree branches, last row).

Output: [very close, close, medium, far, very far]

Depth estimation: Dense Prediction Transformer(DPT)

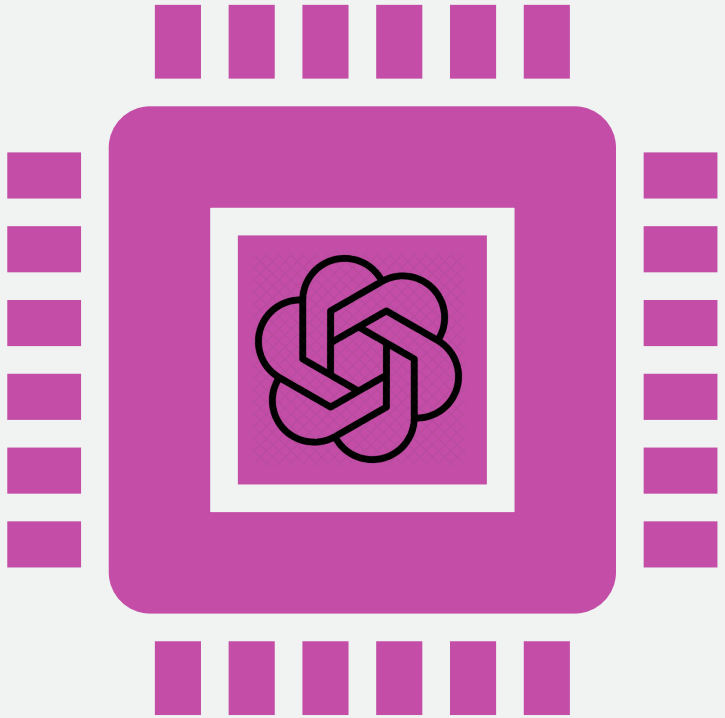
- [2021SOTA]
- Transformer based monocular depth estimation
- can generate image-size NumPy array indicating the depth value of each pixel
- Based on it, we obtained the estimated distance from dashcam to object



Output: [very close, close, medium, far, very far]

LLM: gpt3.5_turbo_16k

- long memory, can analyze the road scene with appropriate prompting
- output can be seen here (0728 version):
- [Summer_Research_2023/Blind_LLM_Guide_Project/Chatbot/]
- There was no long-context or RAG for long video.



Visionary Co-Driver

SYSTEM DESIGN

So, how do we connect those modal together into a pipeline?

How to use LLM locate potential risk?

--- Output format

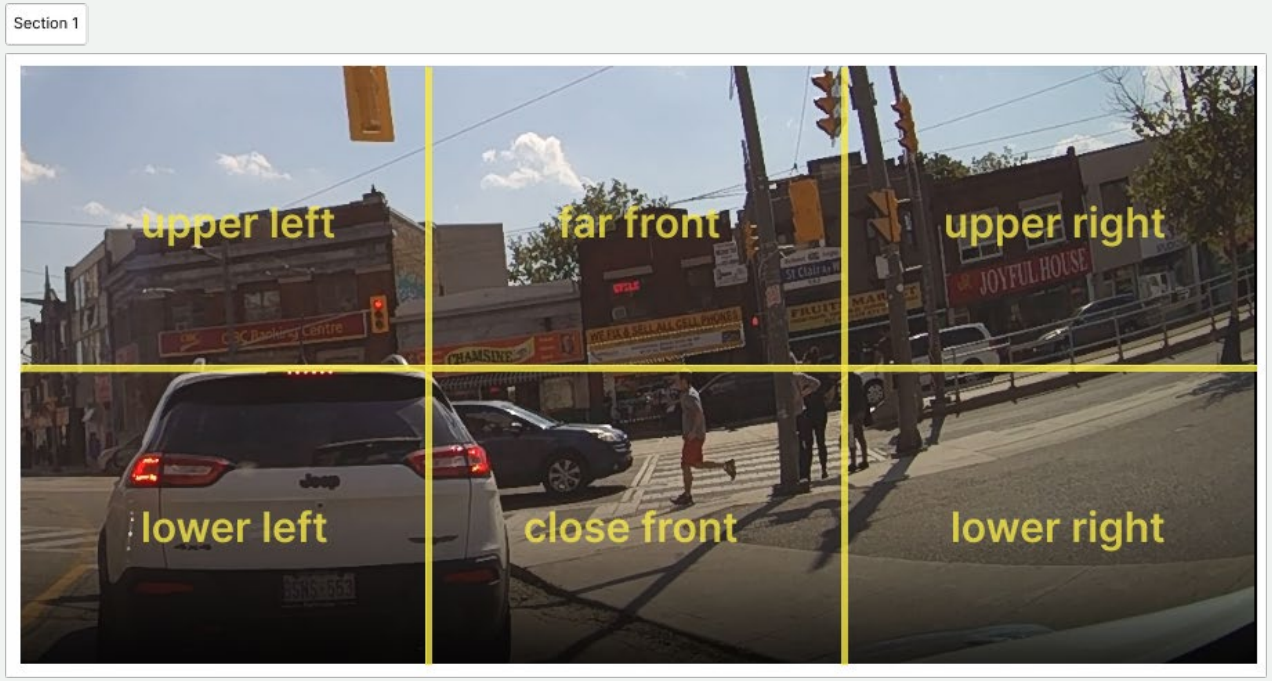
According to early test on LLM’ s capability, it can evaluate the danger level and potential risk within the current driving scenario:

Pic: Zero-shot justification of gpt3.5 turbo

How to use LLM locate potential risk?

--- Output format

According to early test on LLM’ s capability, it can evaluate the danger level and potential risk within the current driving scenario:

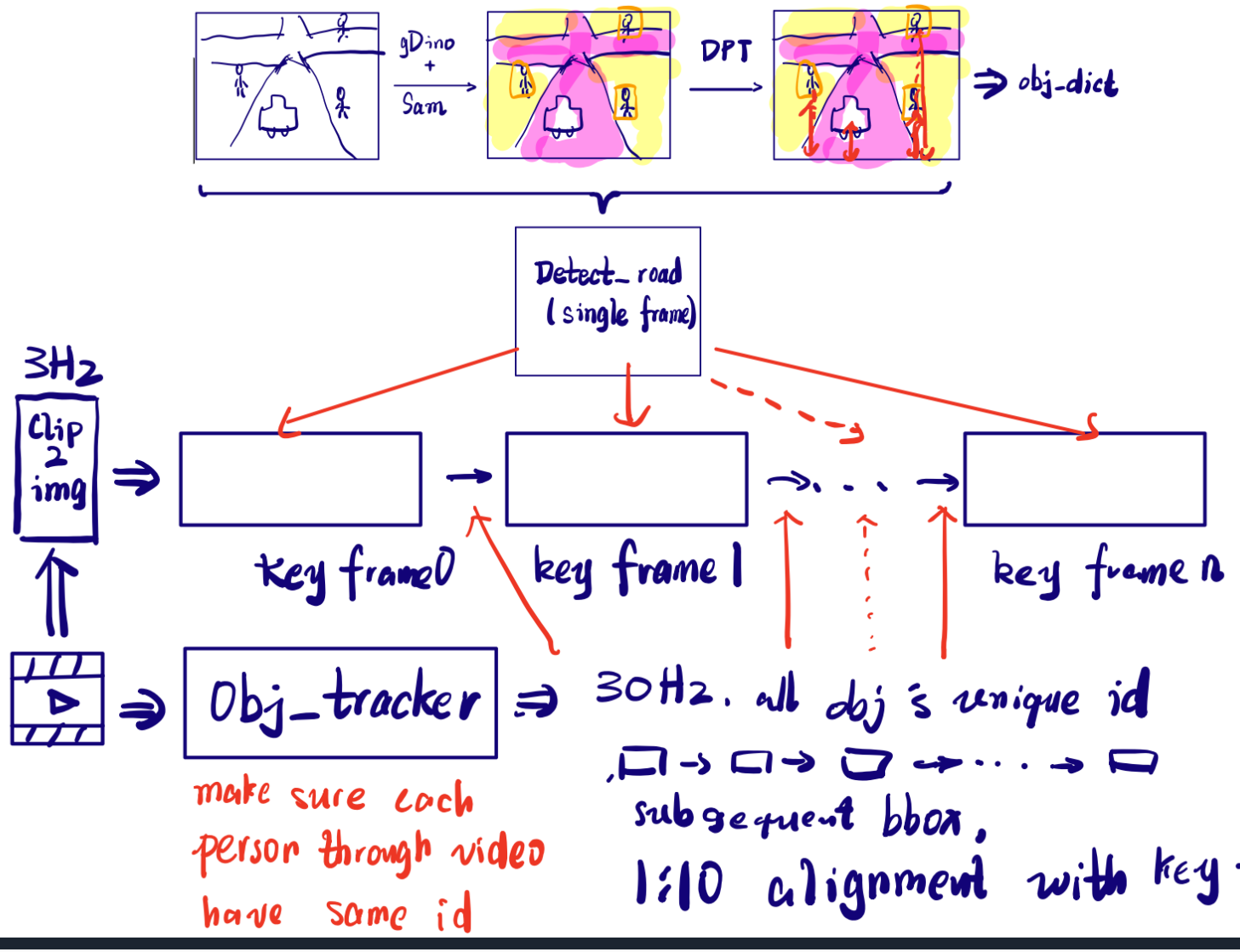


Pixel level classification, divide into 6 aspects

Vision2text

Vision information

Obtain process



How to use LLM locate potential risk?

--- Output format

Furthermore, how can we demonstrate it in:

- Vision: ARHUD
- Unity enhanced Video

Our solution: 关键帧+混合帧分析

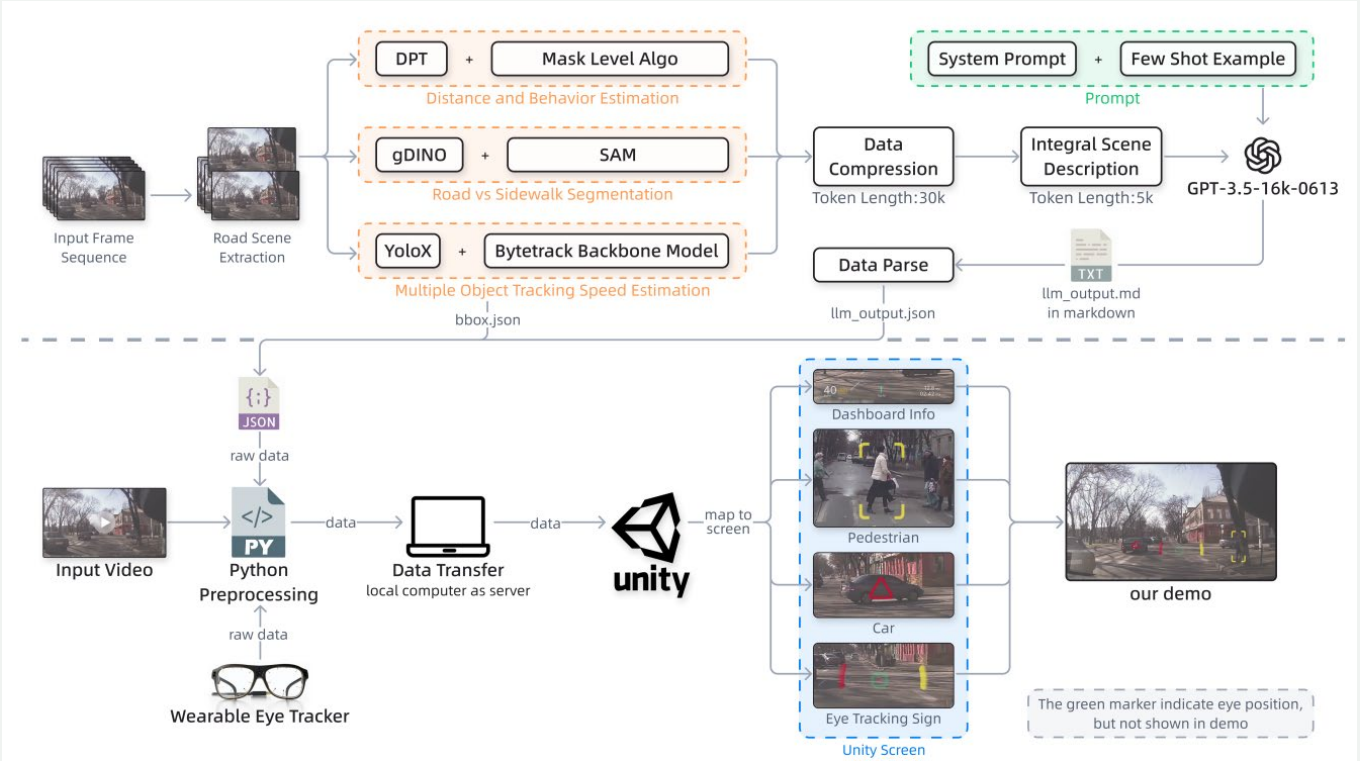


Fig. 2. System design of visionary co-driver

Visionary Co-Driver

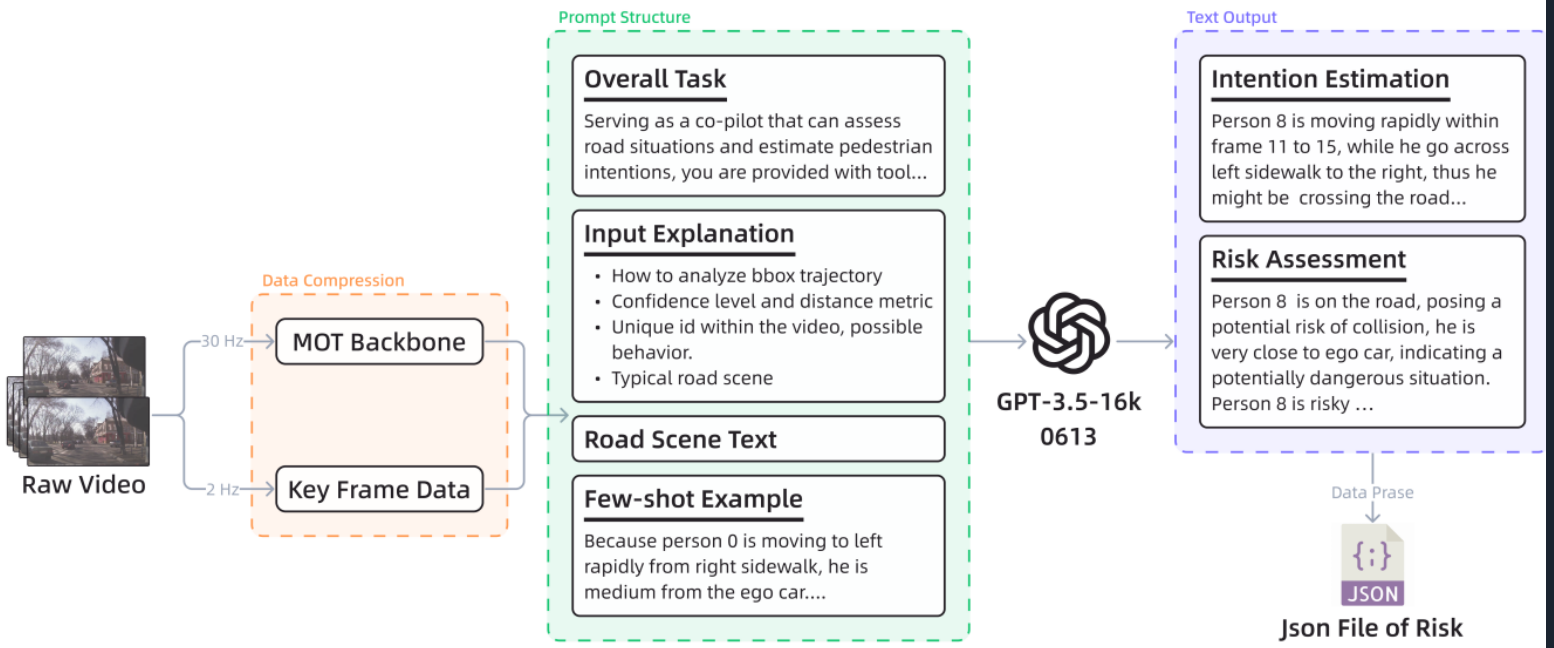


Fig. 8. Reasoning process of Visionary Co-Driver

VISION2TEXT

Text generation process

Vision2text

Output format

JSON/dict form for readability

数据包要求

郑蕲 王杰 | 8月5日创建

- 整体数据
  - 视频的分辨率
  - 视频长度（帧数）
- 每一帧（或者每一个处理的关键帧）的数据
  - 风险的数量
  - 场景描述
  - 每一个风险的几何中心点？（[x, y]）
  - 每一个风险的bbox（xyxy）（或者单独人的风险）
  - 每一个风险的等级（直接风险，潜在高风险，潜在低风险）

典型案例：
Video file: video_0194.mp4

Resolution: 1920x1080

Number of frames: 540

Frames per second: 30.0

Duration (seconds): 18.0

- 场景描述

Info_0194_mask.zip
48.96MB

Info_0194_box.zip
111.26MB

```
91            "motorcycle": {},
92            "bus": {},
93            "truck": {}
94        }, Undetected objects are None
95    Frame "0001": {
96            "person": {
97    Unique Id       "1": [
98    across whole video
99    Bounding box (xyxy)
100                       604.81,
101                       637.4,
102                       641.51,
103                       733.1999999999999
104                   ],
105                   "2": [
106                       1432.43,
107                       650.07,
108                       1462.8700000000001,
109                       712.71
110                   ],
111                   "3": [
112                       39.41,
113                       631.42,
114                       74.97999999999999,
115                       714.9699999999999
116                   ]
117           },
118           "car": {
```

ARHUD

Augmented Reality Heads-Up Display (AR-HUD) interfaces

Fig. 9. Risk sign form example

ARHUD

Augmented Reality Heads-Up Display (AR-HUD) interfaces

Visionary Co-Driver

ARHUD

Augmented Reality Heads-Up Display (AR-HUD) interfaces

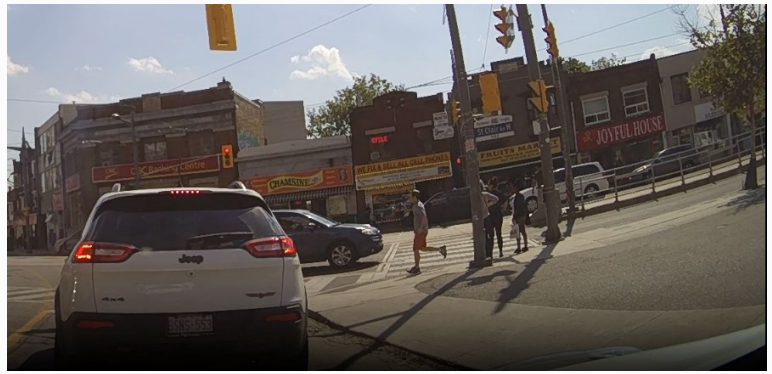
Fig. 11. HUD design presentation

Another basic problem: how to test the model?

It occupies a lot of time to do simulation and real-life testing
So, we choose video,
- monocular
- ground truth justification
- many pedestrian

JAAD dataset

--- the video dataset for pedestrian intention prediction



What is JAAD? Joint Attention in Autonomous Driving (JAAD) Dataset.

Purpose: Developed to study pedestrian behavior, interactions, and visual cues in the context of autonomous driving.

Content:

- **Videos:** Over 240 high-resolution video clips from different urban locations.
- **Annotations:** Includes detailed annotations for pedestrians, such as age, gender, actions, and signals.
- **Scene Information:** Traffic conditions, weather, lighting, etc.
- **Attention and Intention:** Captures the complex interplay of pedestrian attention and intention.

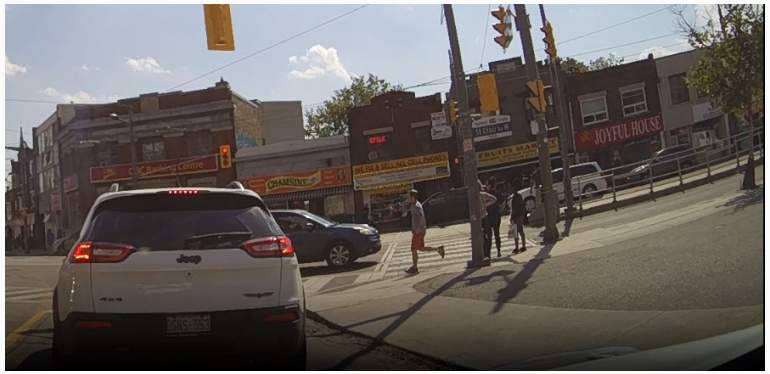
Applications: Crucial for training models to recognize pedestrian actions, enhance traffic safety, and facilitate the development of advanced driver-assistance systems (ADAS).

Challenges: Addresses the need for fine-grained analysis, such as recognizing subtle cues from pedestrians.

Website: Link to JAAD Dataset

JAAD dataset

--- the video dataset for pedestrian intention prediction



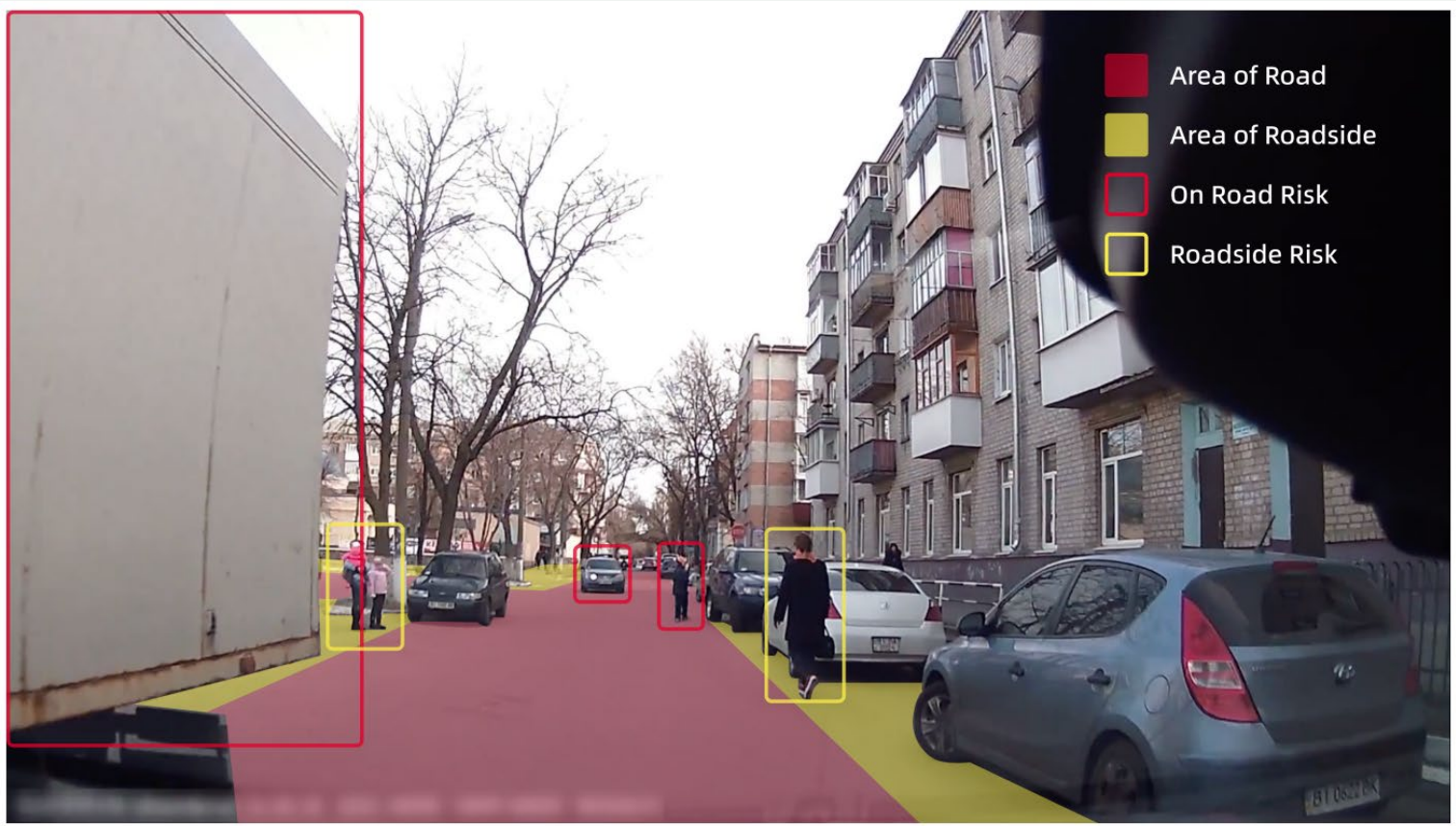


Fig. 1. A typical road scene driver perceives in a cross road, with color-masked separation of 'on road' and 'roadside' area. On the left and far front, a truck, a car and a standing pedestrian is labeled with 'On Road Risk'. On the left and right, several standing pedestrian and a walking pedestrian is label with 'Roadside Risk'.

Visionary Co-Driver

USER STUDY

How do we evaluate the system in a Human-Computer Interaction Experiment?

User Study Procedure (by my teammate)

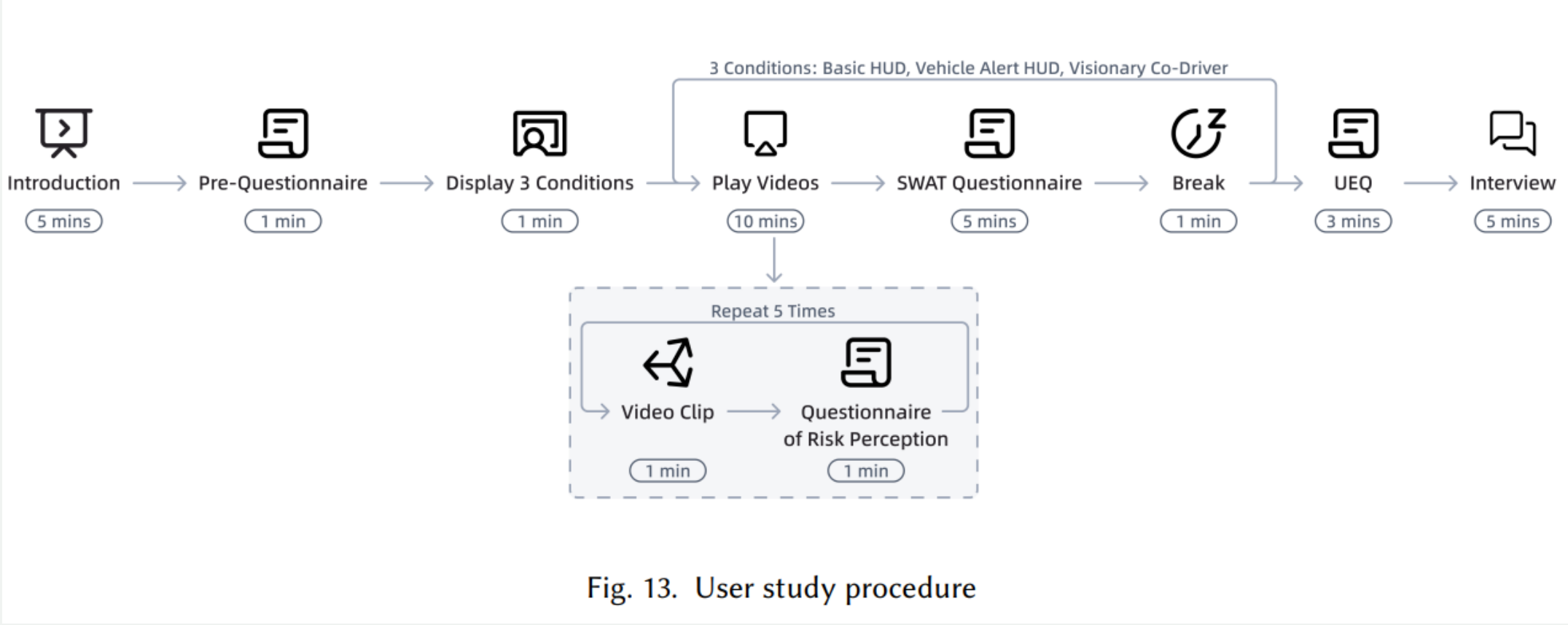


Fig. 13. User study procedure

Keywords: 非注意力盲视，分心，认知负荷，SWAT, UEQ

Visionary Co-Driver

FAILURE & SUCCESS

What did I learn from this summer research projects?

Visionary Co-Driver

Failure Case 1
System Delay

The connected model system is too slow.

Real time consideration is important,

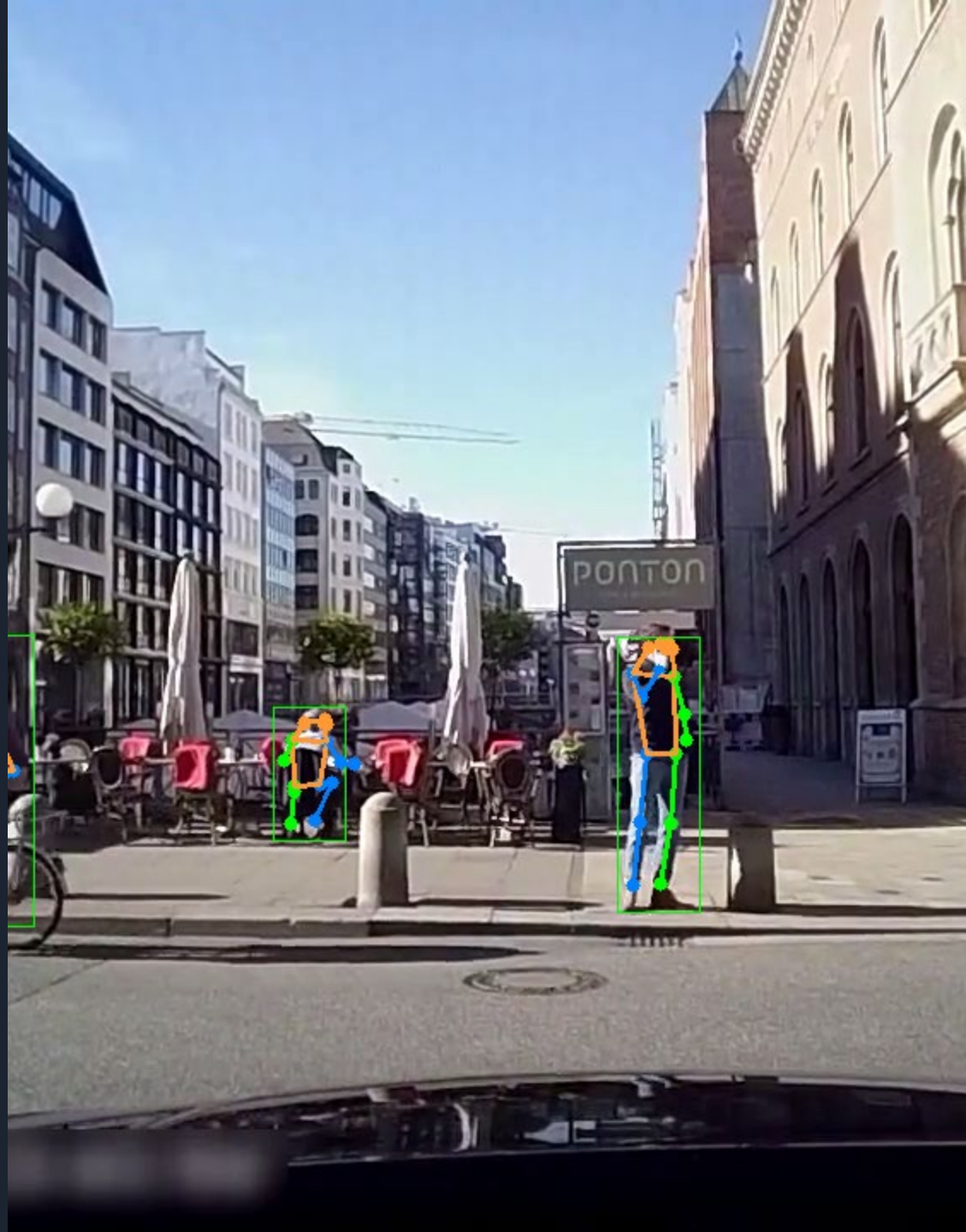
But may be not so important

Solution:

Preprocessing for the ARHUD

Failure Case 2
Pose Estimation

I want to involve Pose direction for further explanation and description, but It is complex and quite hard to analyze the pose bone (You may still require a neural network to decide which way the man is)



Visionary Co-Driver

QA: (What is the object on the counter corner?, microwave)

QA: (How many doors are open?, 1)

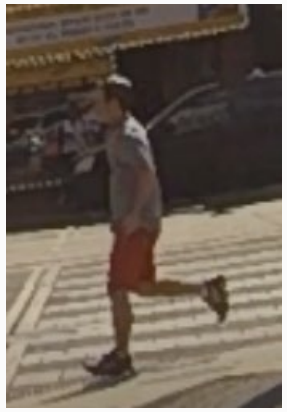
QA: (Where is the oven?, on the right side of refrigerator)

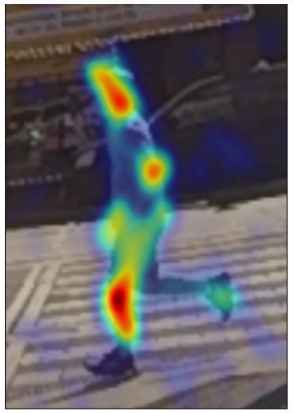
Failure Case 3: VQA(Visual Question Answering)

Given a picture and text question(prompt), the model can give appropriate feed back

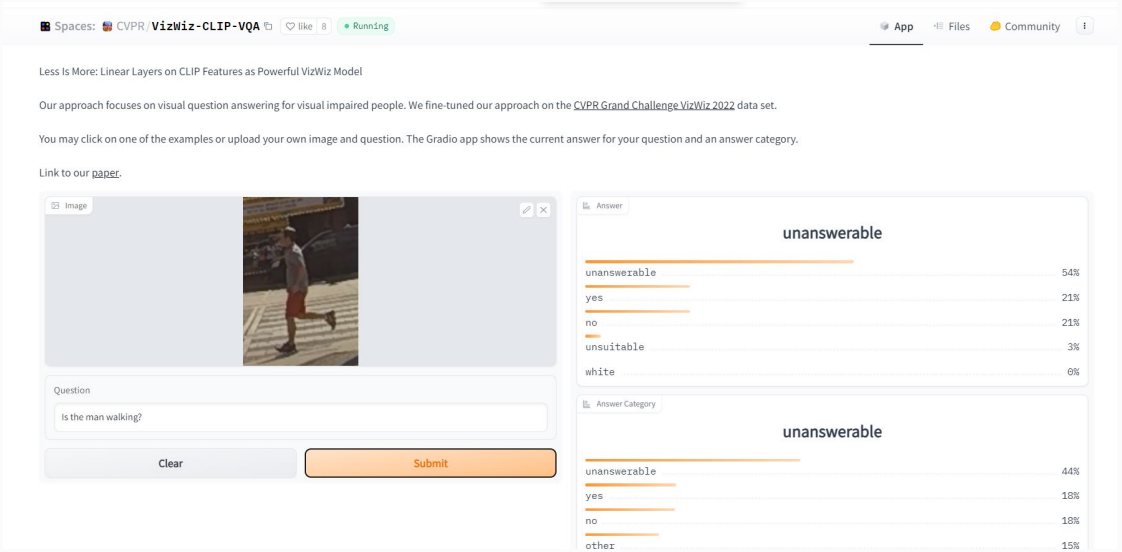
- Natural approach of considering the vision2text task
- however, it don’ t work for now, too stupid. There is still not practical modal for common sense justification.

Visionary Co-Driver





human



Failure Case 3: VQA(Visual Question Answering)

- Can’t recognize color the person is wearing
- Can’t distinguish left and right
- Can’t distinguish young or old
- It is too unstable(50% success rate), so we don’t use it

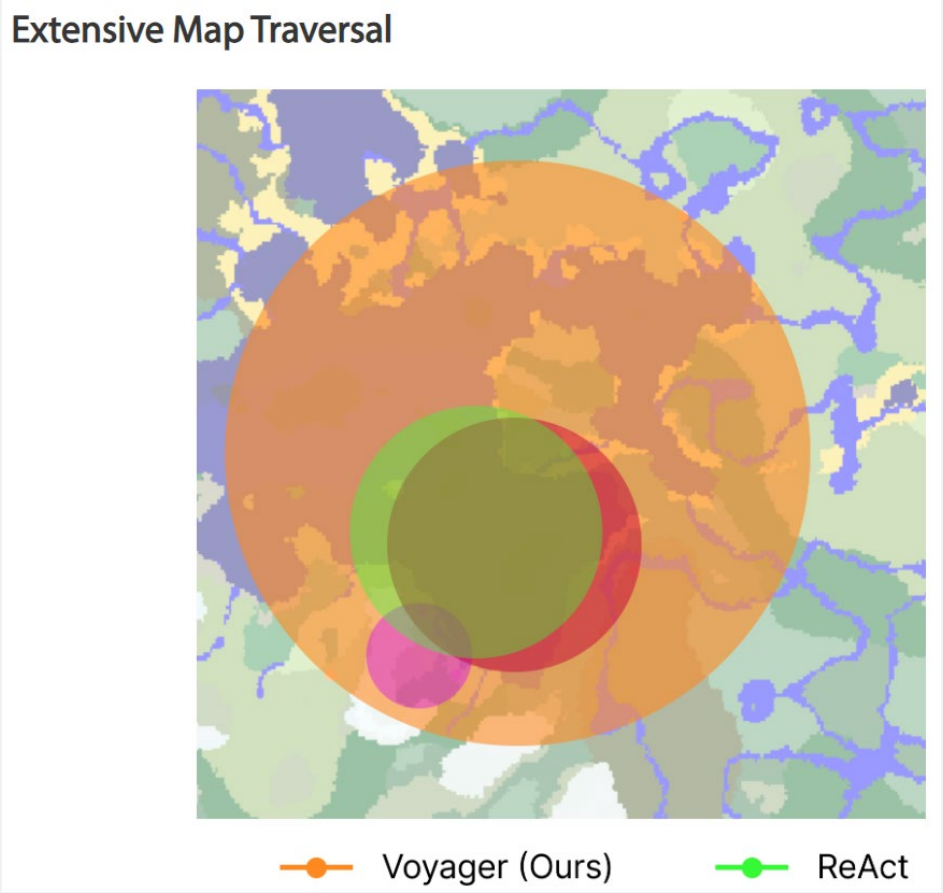
Visionary Co-Driver

Possible IDEA: LLMs as Agent

Intuition: Let GPT itself decide when to function call other api, obtain more information from the current screen.

Because

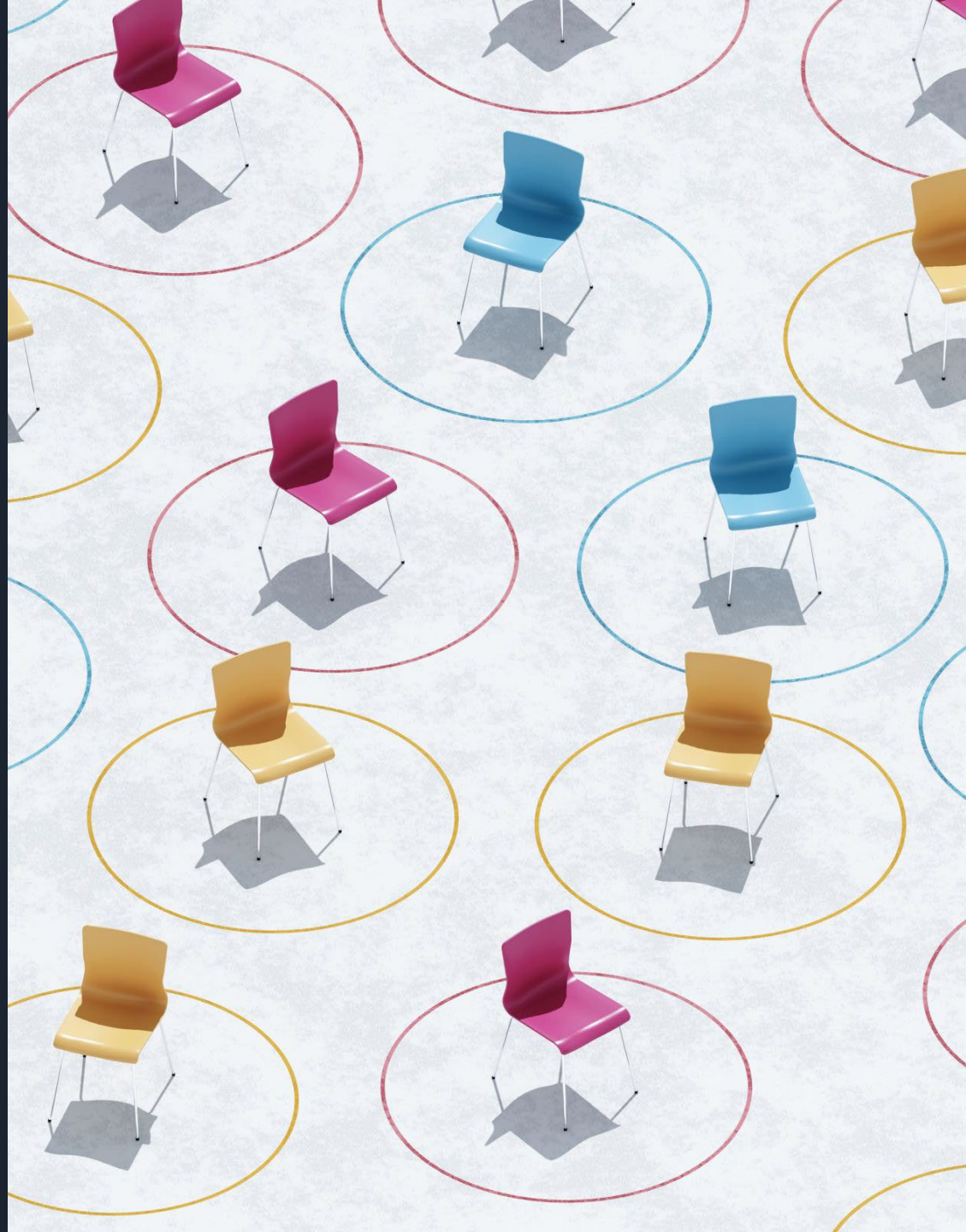
- the VQA performance can be enhanced in Multiple Object scenario.
- This sounds like a more ‘human-like’ behavior
- Other Agent: Voyager, AutoGPT…



Failure Case 4

Think too much about the COT

Being a big fan of AI agent like Voyager and code-interpreter, I spent a long time on this huge task. But maybe there exists a gap between personal willing and his ability. Those fancy idea requires a large team and a long time to succeed.



Blind, But Visionary Co-driver

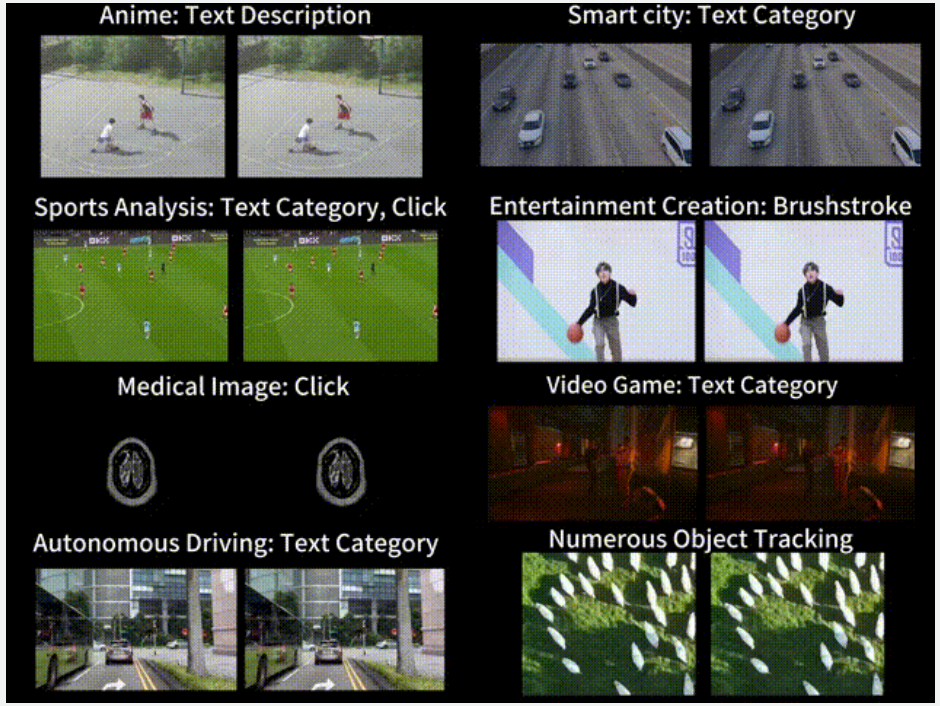
LLM不应该只是被动的收集信息，他应该主动调用API， 去探测前面的路口状况，让他作为一个Agent去执行分析任务

这样就可以自动化的收集信息，分析数据了

比方说可以写点segment 的API， 让他可以把一个image的interesting area/ box 切割出来，再图生文，用VQA 或者再调用现有api，输出更多的text info

Possible Combination: Segment and Track Anything

- [SOTA Mask MOT]
- ZJU Work
- Based on SAM and AOT(Associating Objects with Transformers, also ZJU work)
- tracking and segmenting any objects in videos, either automatically or interactively.
- Promptable -> gDINO for text
- What a pity: Not use this earlier



Visionary Co-Driver

Our Success

Visual inputs: VGA charger
Sample 1 of 7

Next sample

User

What is funny about this image? Describe it panel by panel.

Source: hmmm (Reddit)

Why is our work more valuable even after visual GPT4 come out?

Vision是内含知识的，这种逻辑推理是显式的（implicit），我们的方案可解释性更高，而且算力需求更小（visual 如CLIP等动辄10~20s一张图）

驾驶场景的优先度，场景拆分，深入描述

All need further consideration, visual GPT4 can’t handle all the case as well

Some random thought: those might be useful hint

1. Be careful of the translation! Once I misunderstand '人行道' to be 'sideroad', which greatly lower the performance of gDINO! [prompt engineering]

    - in fact: sidewalk / pavement

2. Treat ChatGPT as your project co-pilot with Custom Instruction, while be careful to its local optimum!
3. Engaging in engineering requires reading more research papers, as it makes the technical details more concise and valuable. GPT's knowledge base capability is limited, often broad but not insightful enough. If necessary, let GPT help generate keywords, and then we can search on Google Scholar ourselves.
4. Also, it really helps to watch tutorial sometimes. I should review Andrew Ng's ChatGPT course from time to time.

Some random thought: those might be useful hint

5. The research ability equals engineering ability. A practical shitty draft is better than any dreamy draft. '**Talk is cheap, show me your code.**' The way I am interested in exploring new insight is coding.

6. Meanwhile, I found it very helpful to write down the messy idea in English when I felt bad. Maybe it is because of mother tongue shame I suddenly escape from my mental local minimum.

7. Trust and collaborate with teammates on different parts of the project, like HUD design.

8. Seek Expert Opinions: Consider engaging with academics and professionals in the field for insight and potential collaboration (getting push from them)!

Visionary Co-Driver

THE END

That’s all, thank you for listening!