MME-RealWorld: Could Your Multimodal LLM Challenge High-Resolution Real-World Scenarios that are Difficult for Humans?

Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, Liang Wang, Rong Jin, Tieniu Tan

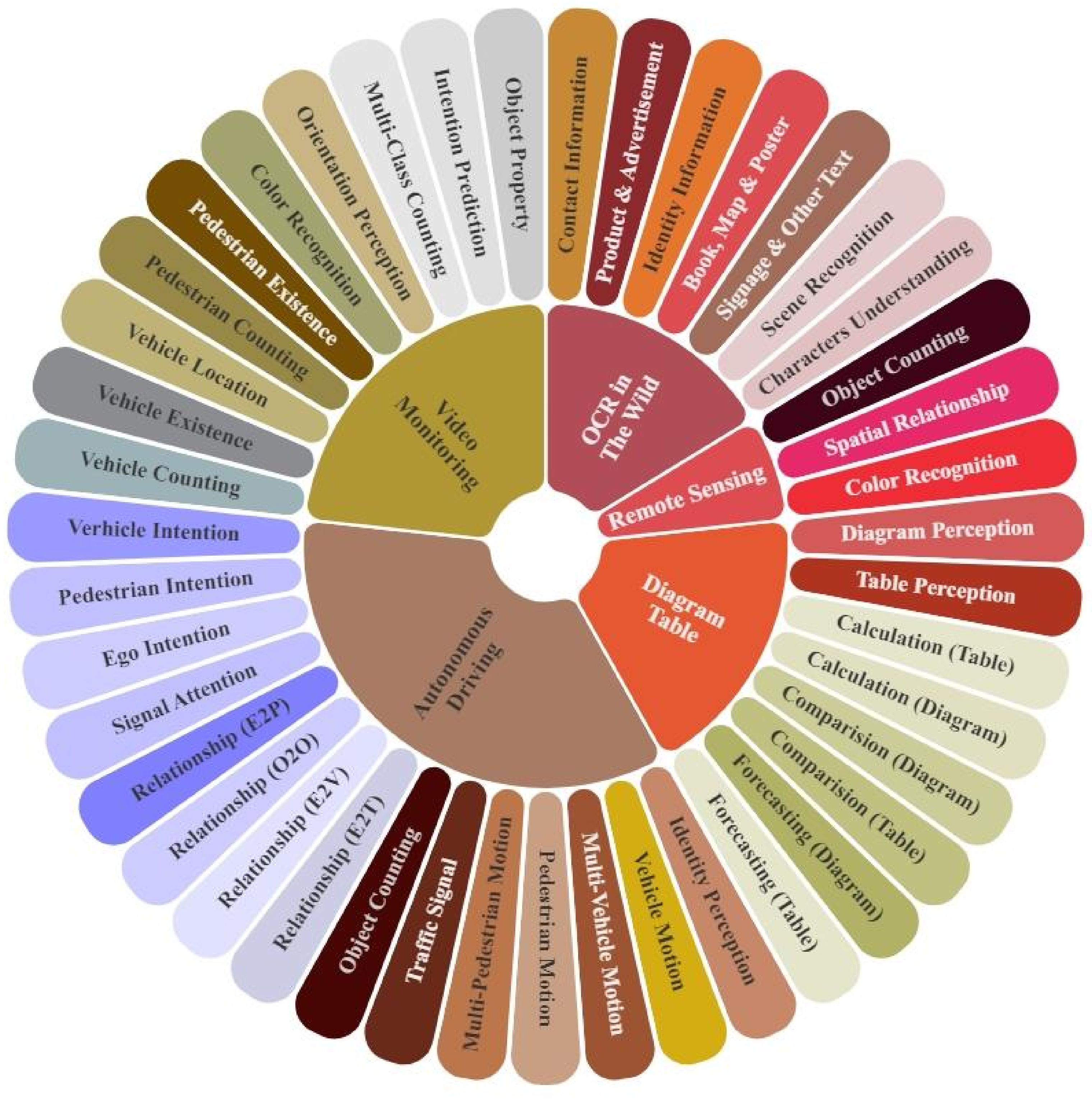
TL;DR: Existing benchmarks for MLLMs face limitations such as small data scale, model-based annotations, and insufficient task difficulty, hindering their ability to measure real-world challenges. To address these issues, we introduce **MME-RealWorld**, a **large-scale**, **manually annotated** benchmark with **high-resolution** images and complex real-world scenarios, revealing that even advanced models like GPT-4o and Gemini 1.5 Pro struggle to achieve 60% accuracy, highlighting the urgent need for improved perception and understanding capabilities.

Main Motivations

- **Data Scale** Many existing benchmarks contain fewer than 10K QA pairs ;
- **Annotation Quality** Some benchmarks, while larger in scale, rely on annotations generated by LLMs or MLLMs, whose annotations are inherently limited by the performance of the used models.
- **Task Difficulty** The top performance of some benchmarks has reached the accuracy of 80%-90%, and the performance margin between advanced MLLMs is narrow.

MME-RealWorld

- **Data Scale** We have manually annotated **29, 429** QA pairs focused on real-world scenarios.
- **Data Quality**
- **Resolution:** The images have an average resolution of 2, 000×1, 500, containing rich image details.
- **Annotation:** All annotations are manually completed.
- **Task Difficulty and Real-World Utility**
  Even the most advanced models have not surpassed **60%** accuracy.
- **MME-RealWorld-CN** We collect 5, 917 QA pairs focusing on Chinese Scenarios, annotated by Chinese volunteers.

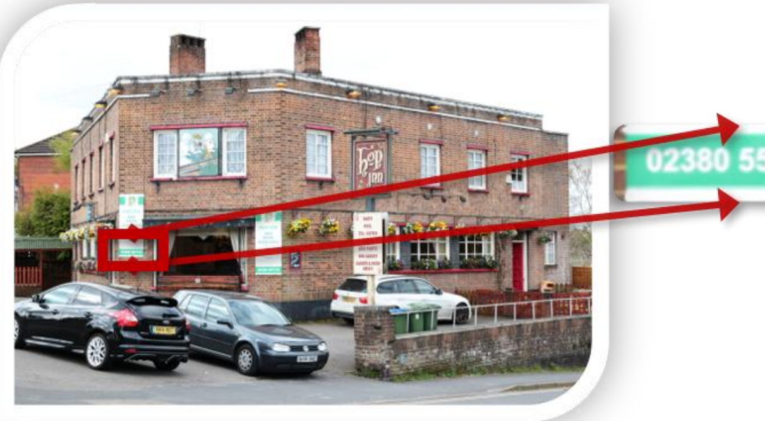


(a) **Real-World Tasks**

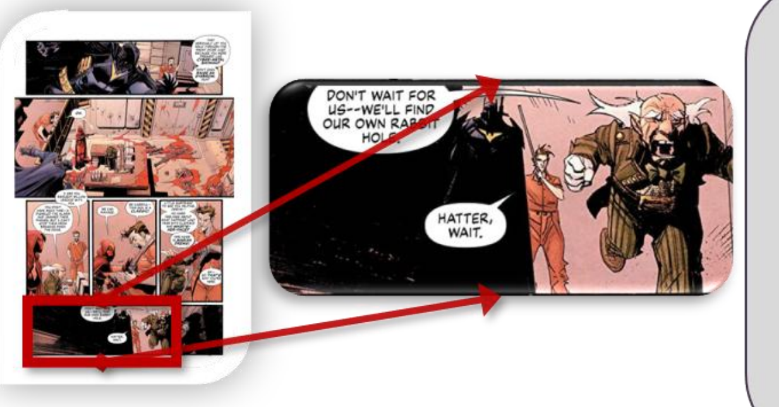
Eng		CN	
Model	**Acc**	**Model**	**Acc**
QwenVL-2	56.5	InternVL-2	55.5
InternVL-2	53.5	QwenVL-2	55.5
Claude 3.5 Sonnet	51.6	InternVL-Chat-V1-5	47.9
InternLM-2.5	50.0	Claude 3.5 Sonnet	47.0
InternVL-Chat-V1-5	49.4	SliME-8B	45.8
Mini-Gemini-34B-HD	45.9	YI-VL-34B	42.0
MiniCPM-V 2.5	45.6	CogVLM2	39.8
GPT-4o	45.2	SliME-13B	38.9
CogVLM2	44.6	GPT-4o	38.8
Cambrian-34B	44.1	Mini-Gemini-34B-HD	38.5
Cambrian-8B	42.7	Monkey	37.2
SliME-8B	39.6	LLaVA-Next-8B	36.5
Gemini-1.5-pro	38.2	Cambrian-34B	35.7
GPT-4o-mini	36.4	Mini-Gemini-7B-HD	34.9
Monkey	35.3	*InternLM-2.5*	33.9
mPLUG-DocOwl	32.7	Cambrian-8B	33.6
DeepSeek-VL	32.4	LLaVA-Next-72B	30.6
SliME-13B	31.7	mPLUG-DocOwl	28.3
YI-VL-34B	31.0	Gemini-1.5-pro	28.1
Mini-Gemini-7B-HD	30.3	MiniCPM-V 2.5	27.9
LLaVA-Next-8B	30.2	DeepSeek-VL	27.6
LLaVA-Next-72B	28.7	TextMonkey	26.4
LLaVA1.5-13B	28.0	GPT-4o-mini	25.9
ShareGPT4V-13B	27.8	ShareGPT4V-13B	25.9

(b) **Leaderboard**

OCR in the Wild

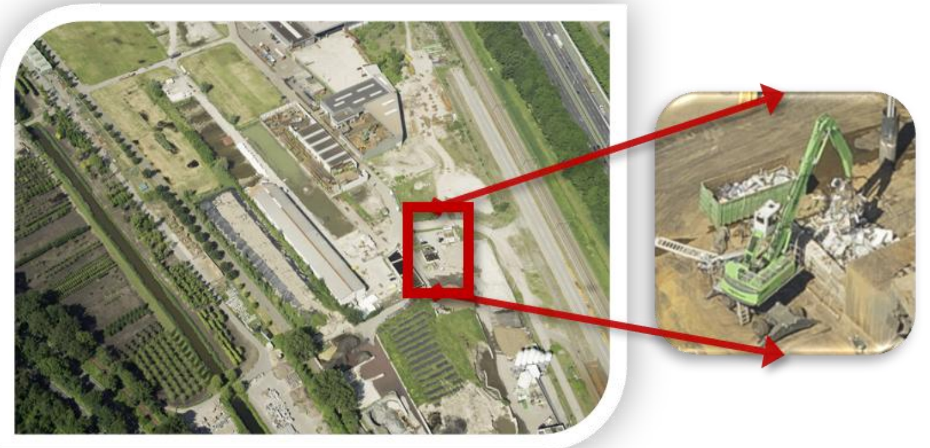


What is the phone number of HOP INN?
(A) 02380 667723
(B) 02880 557722
(C) 02025 557753
(D) 02380 557723
(E) The image does not feature the number.

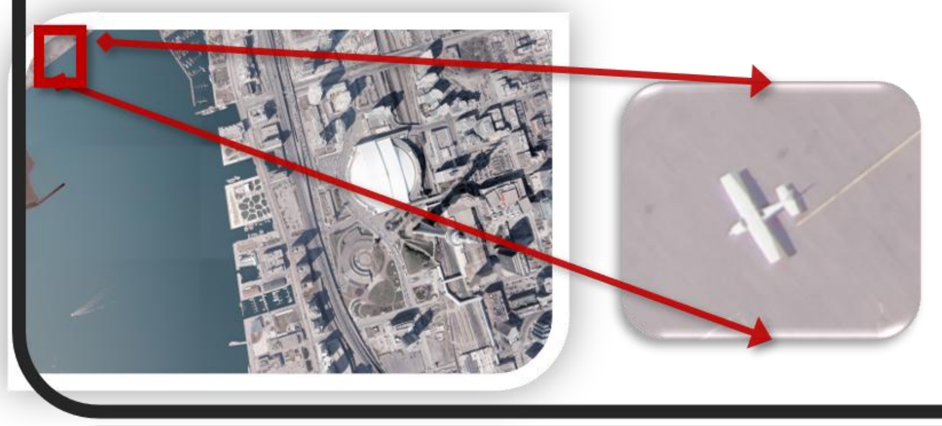


What's the content of the old man speaking in the frame on the fourth row in the comic page?
(A) DON'T WAIT FOR US. WE'LL FIND OUR OWN RABBIT HOLE.
(B) I SEE YOU BROUGHT WILLOW LIFGOOD WITH YOU.
(C) HER NAME IS MARIAN DREWS!
(D) WE CAN MANAGE.
(E) The image does not feature the content.

Remote Sensing

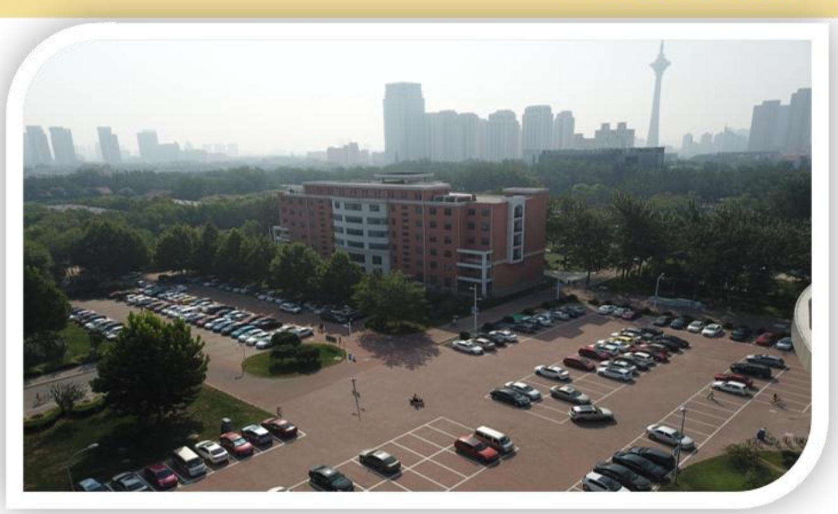


What is the color of the excavator in the middle area of the picture?
(A) Yellow.
(B) Red.
(C) Black.
(D) Green.
(E) The image does not feature the color.

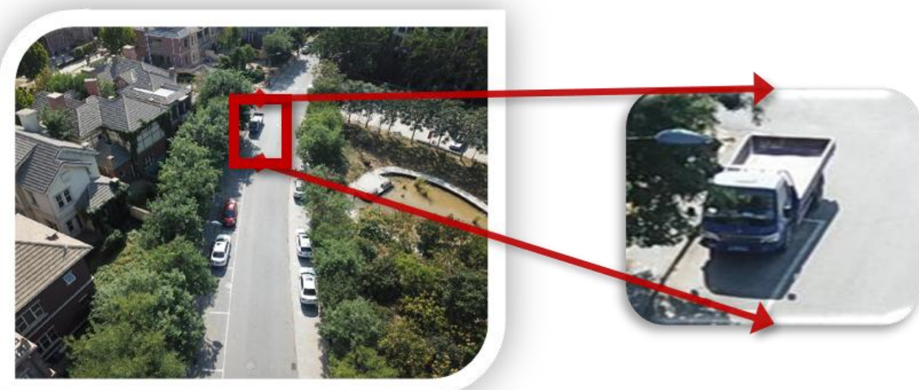


How many aircraft are there in the picture?
(A) 1
(B) 2
(C) 3
(D) 5
(E) 0

Video Monitoring

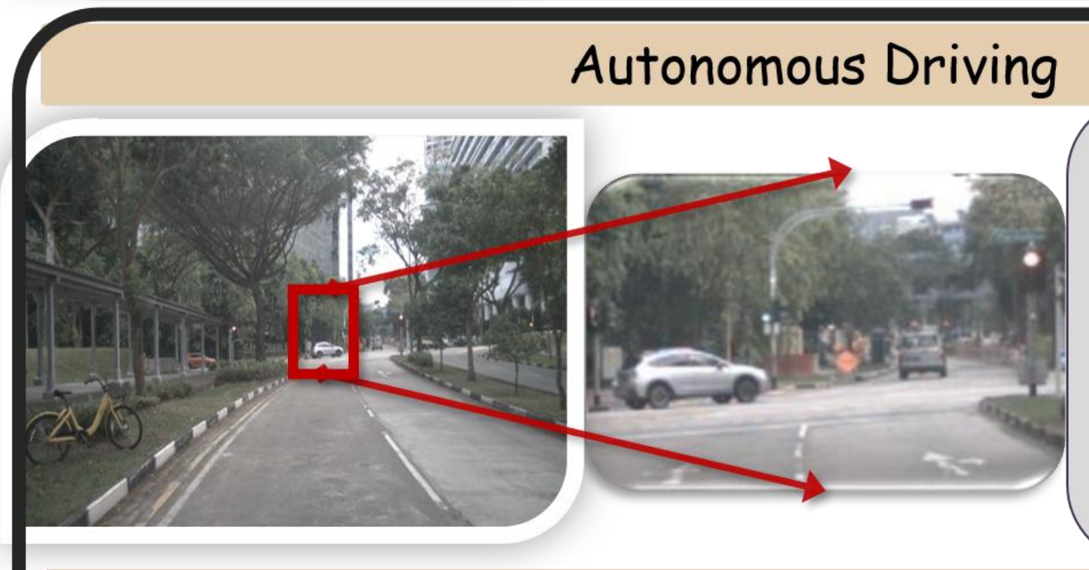


What is the total number of motors and cars in the image?
(A) 26
(B) 80
(C) 133
(D) 92
(E) The image does not feature the objects

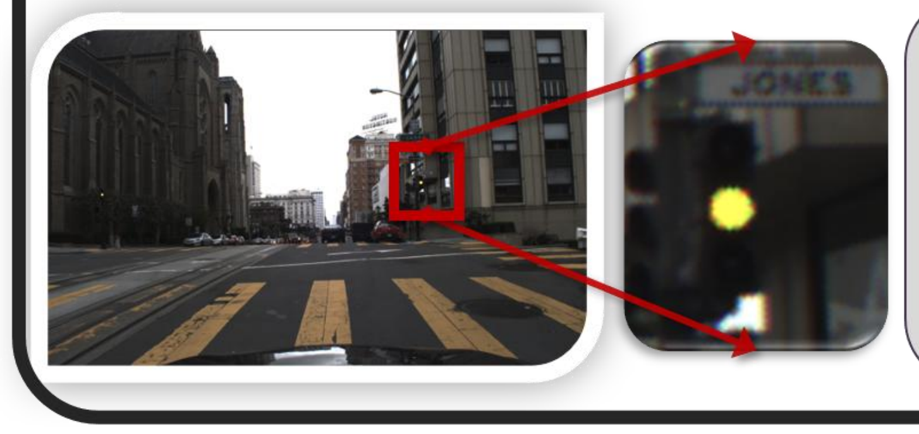


What will the truck do in the image?
(A) Stopping
(B) Keep moving
(C) Turn left
(D) Turn right
(E) The image does not feature the object.

Autonomous Driving

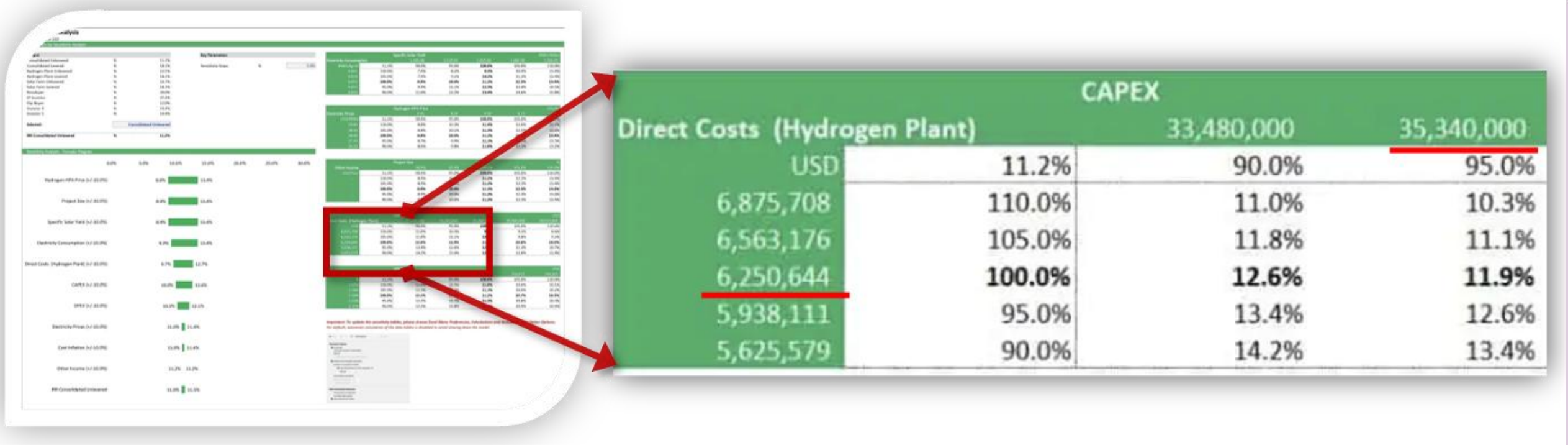


This image shows the front view of the ego car. What is the future state of the white suv in the middle?
(A) Turn right.
(B) Turn left.
(C) Stationary.
(D) Keep going straight.
(E) The image does not feature the object



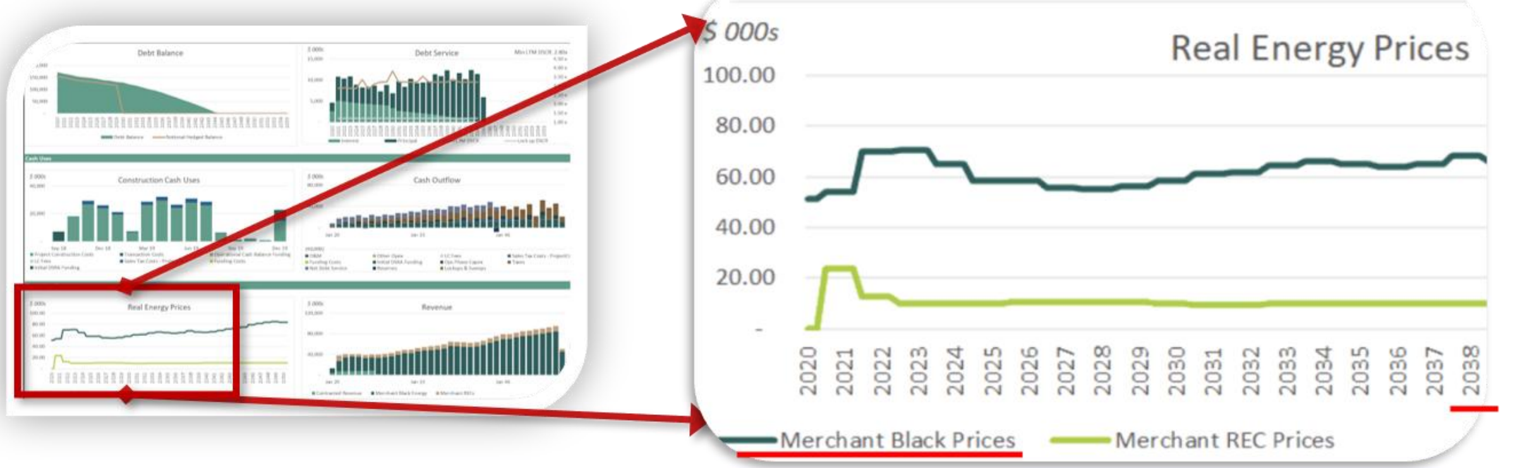
What is the traffic light on the right?
(A) yellow
(B) red
(C) green
(D) changing/off
(E) The image does not feature the traffic light

Diagram and Table



CAPEX			
Direct Costs (Hydrogen Plant)		33,480,000	35,340,000
USD	11.2%	90.0%	95.0%
6,875,708	110.0%	11.0%	10.3%
6,563,176	105.0%	11.8%	11.1%
6,250,644	**100.0%**	**12.6%**	**11.9%**
5,938,111	95.0%	13.4%	12.6%
5,625,579	90.0%	14.2%	13.4%

What's the percentage of CAPEX when direct costs rise to 35340000 if the outcome is at 6250644?
(A) 12.6% (B) 10.3 **(C) 11.9%** (D) 11.1 (E) The image does not feature the DATA.



What is the data of Merchant Black Prices in 2038 in the diagram Real Energy Prices?
(A) 40.00 to 60.00 (B) 20.00 to 40.00 **(C) 60.00 to 80.00** (D) 80.00 to 100.00
(E) The image does not feature the data.

Key Findings:

1) Existing Models Still Lacking in Image Detail Perception.
2) Processing high-resolution images reveals significant disparities in computation efficiency across different models.
3) Error analysis reveals that larger models tend to select safer options, while smaller models favor the first choice.
4) Some open-source models exhibit limited instruction-following capabilities