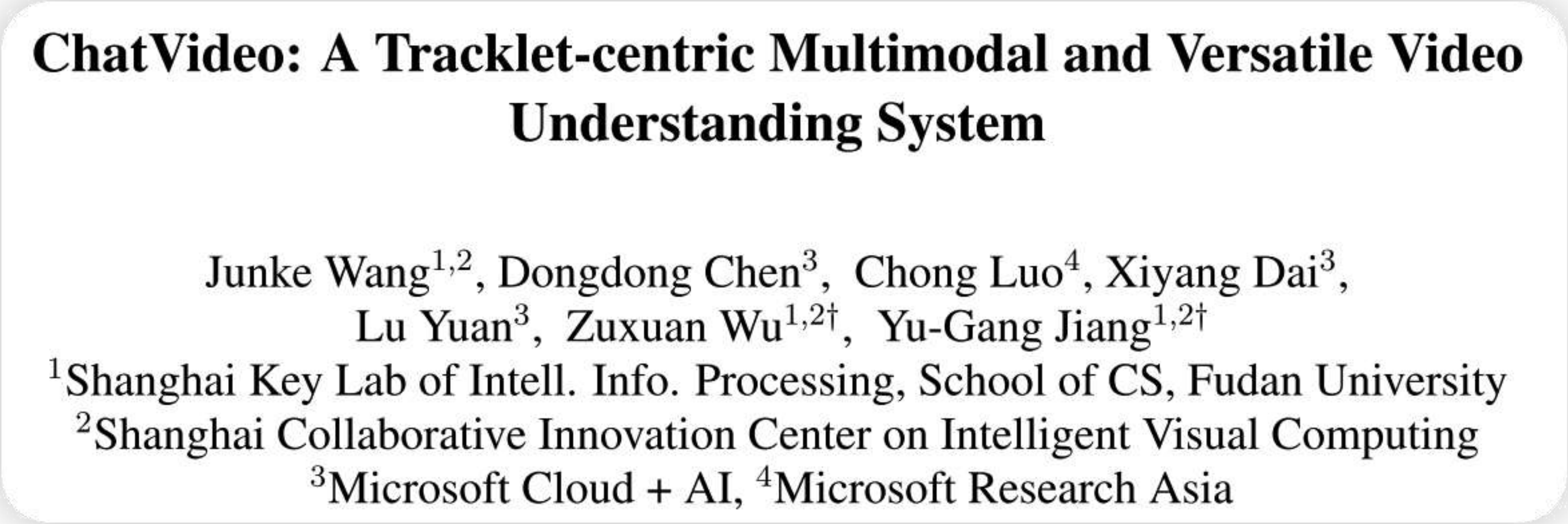
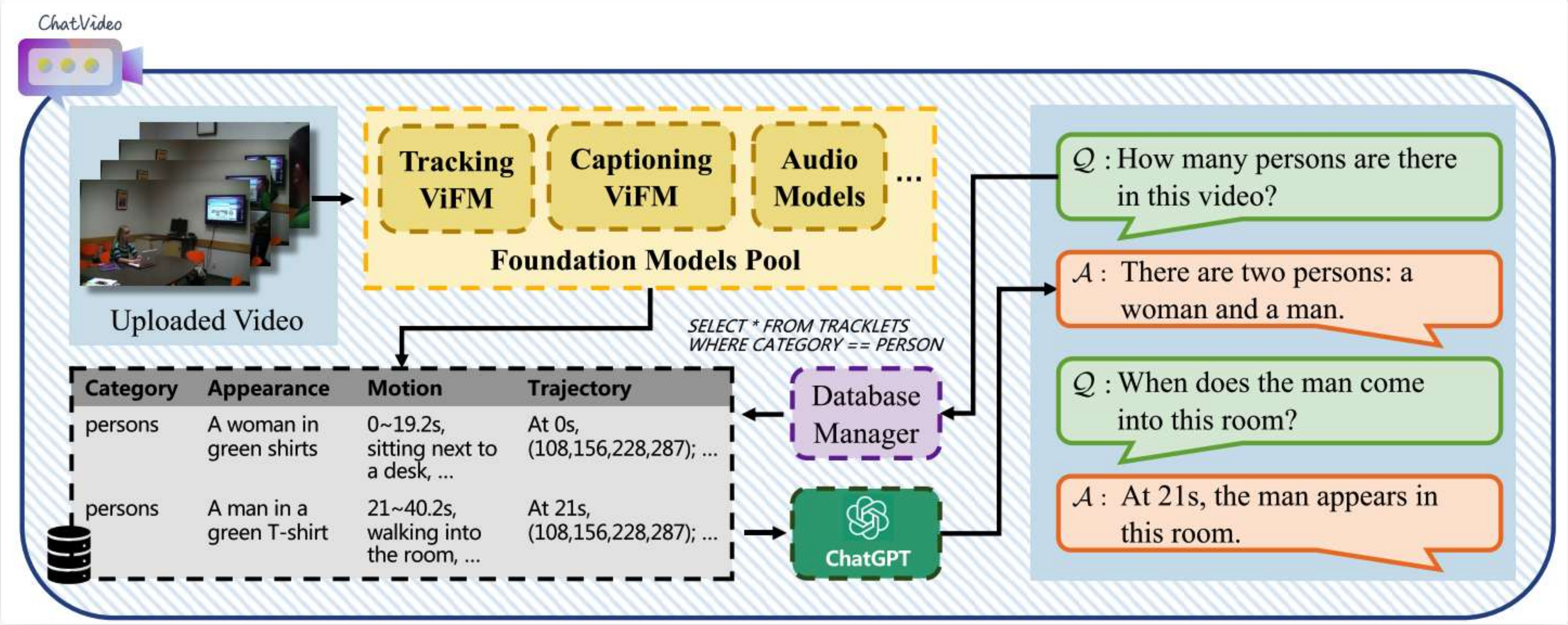


ChatVideo: A Tracklet-centric Multimodal and Versatile Video Understanding System

Junke Wang[1,2], Dongdong Chen[3], Chong Luo[4], Xiyang Dai[3],
Lu Yuan[3], Zuxuan Wu[1,2†], Yu-Gang Jiang[1,2†]
[1]Shanghai Key Lab of Intell. Info. Processing, School of CS, Fudan University
[2]Shanghai Collaborative Innovation Center on Intelligent Visual Computing
[3]Microsoft Cloud + AI, [4]Microsoft Research Asia

Reporter: Yifei Cao
2023.7.13



Category	Appearance	Motion	Trajectory
persons	A woman in green shirts	0~19.2s, sitting next to a desk, ...	At 0s, (108,156,228,287); ...
persons	A man in a green T-shirt	21~40.2s, walking into the room, ...	At 21s, (108,156,228,287); ...

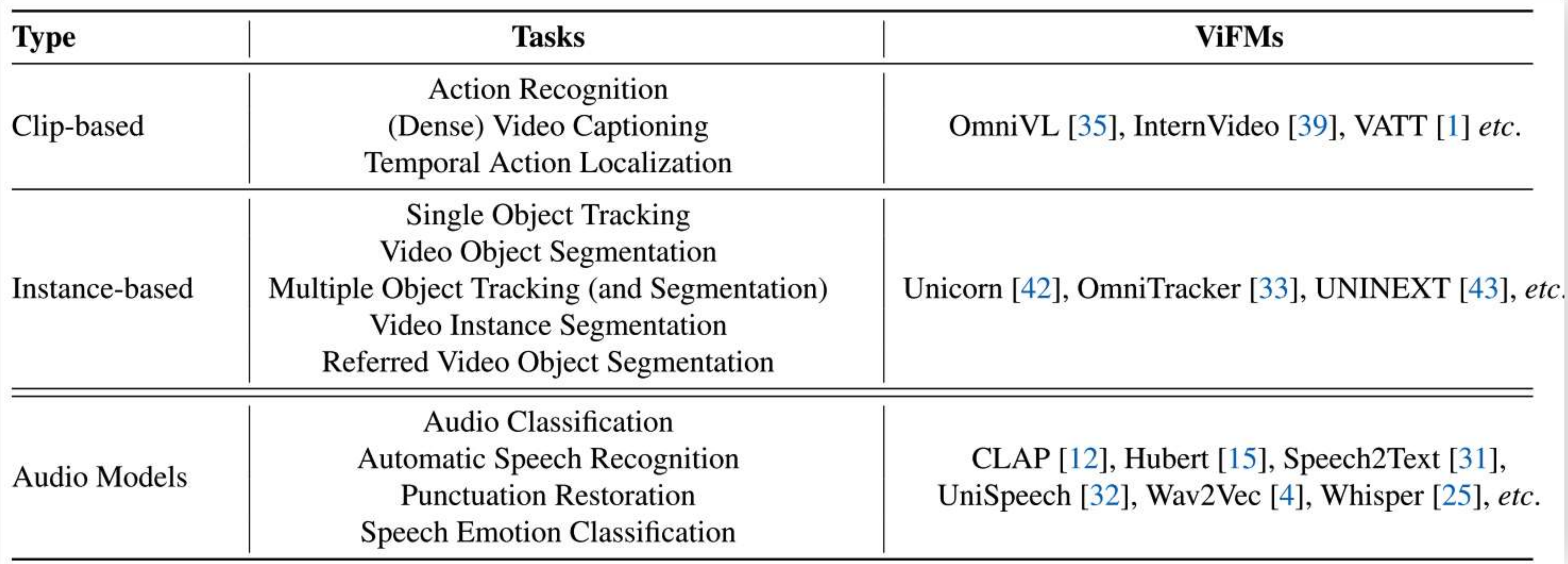
Q : How many persons are there in this video?

A : There are two persons: a woman and a man.

Q : When does the man come into this room?

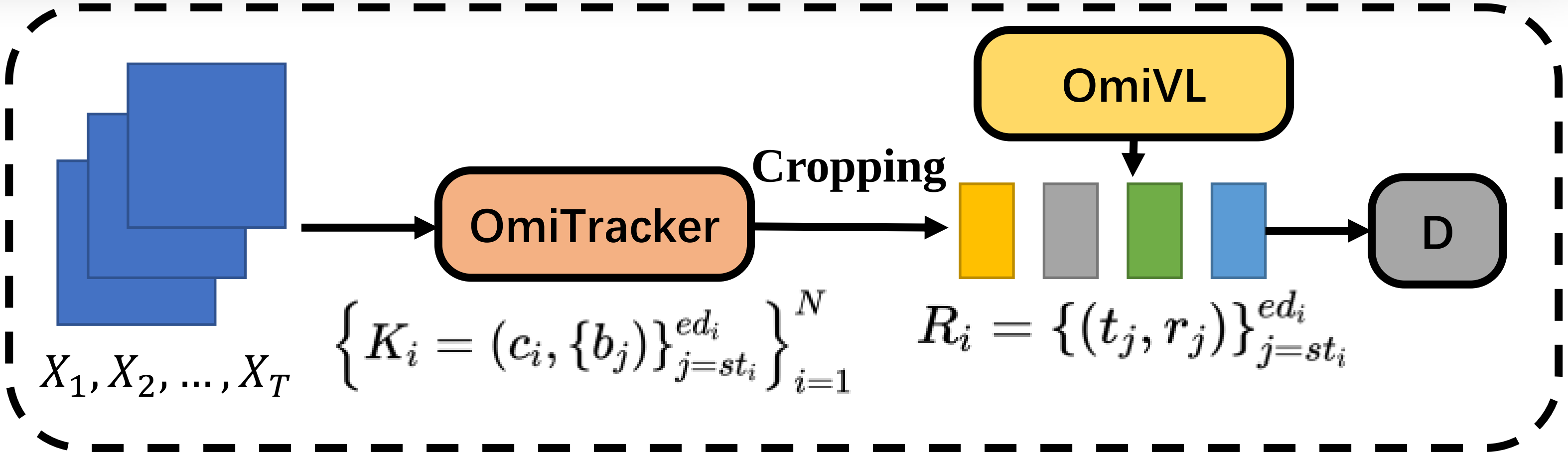
A : At 21s, the man appears in this room.

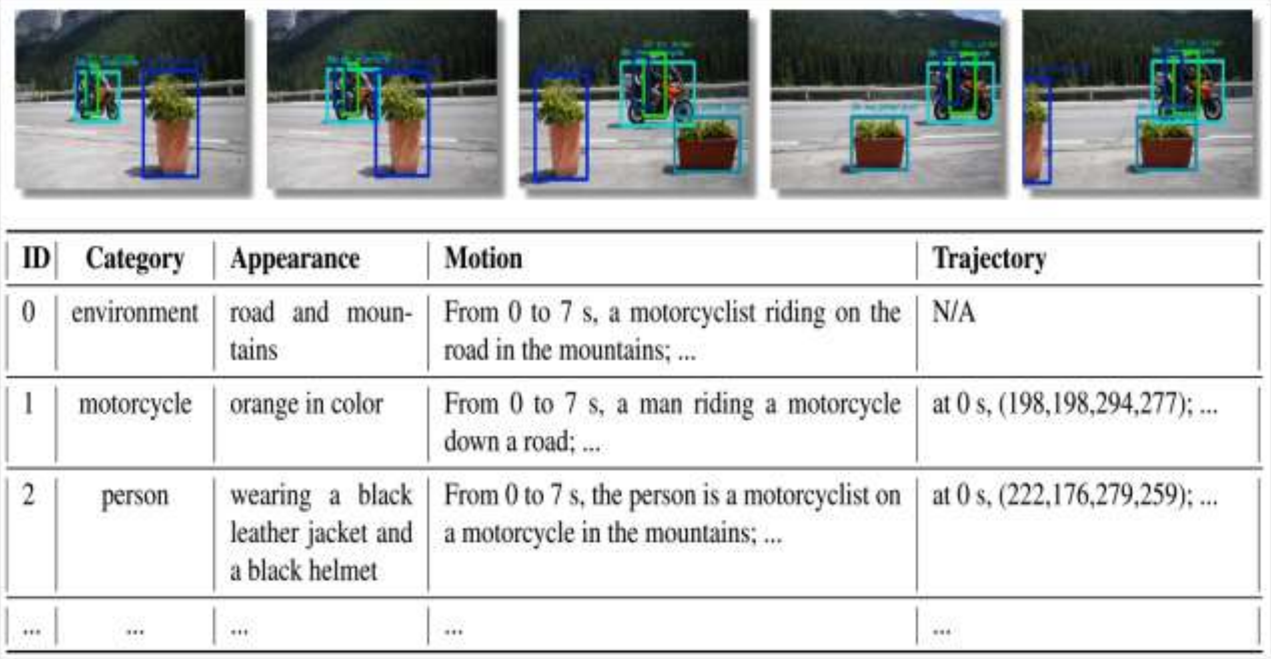
- 为什么visual chatgpt不适合于解决视频任务?
- 关键思路



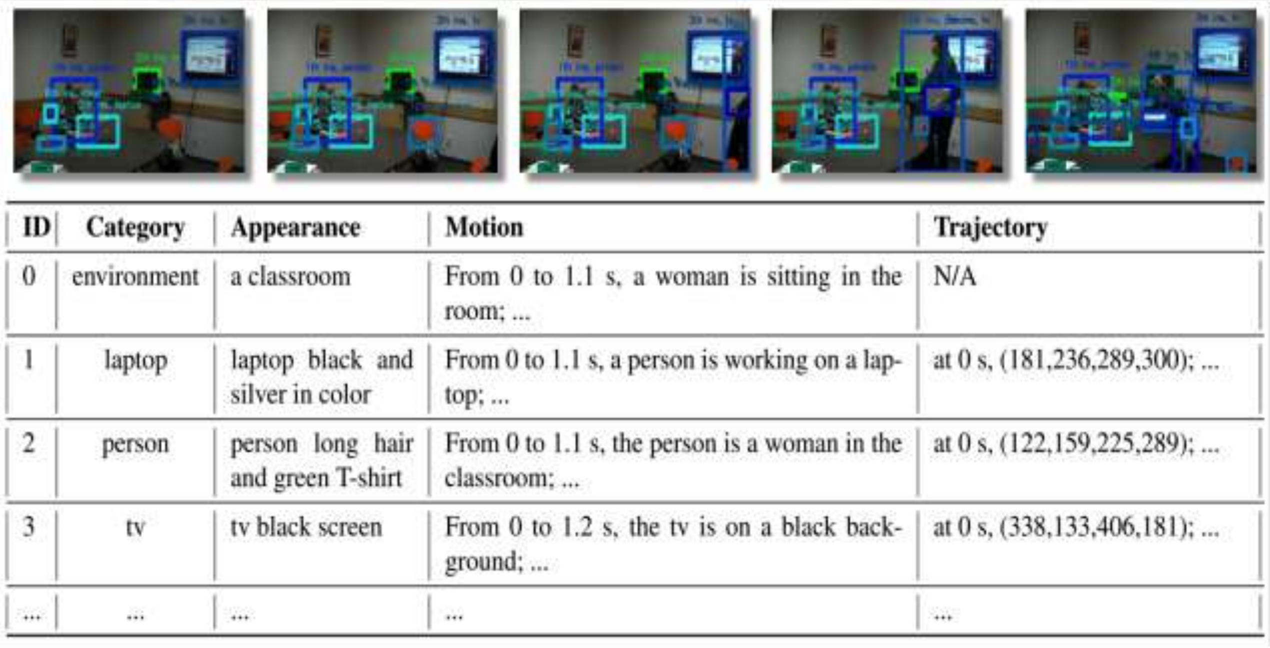
Type	Tasks	ViFMs
Clip-based	Action Recognition (Dense) Video Captioning Temporal Action Localization	OmniVL [35], InternVideo [39], VATT [1] *etc.*
Instance-based	Single Object Tracking Video Object Segmentation Multiple Object Tracking (and Segmentation) Video Instance Segmentation Referred Video Object Segmentation	Unicorn [42], OmniTracker [33], UNINEXT [43], *etc.*
Audio Models	Audio Classification Automatic Speech Recognition Punctuation Restoration Speech Emotion Classification	CLAP [12], Hubert [15], Speech2Text [31], UniSpeech [32], Wav2Vec [4], Whisper [25], *etc.*

Tracklets数据库生成：


$X_1, X_2, \dots, X_T$
OmiTracker
$\left\{K_i = (c_i, \{b_j\}_{j=st_i}^{ed_i}\right\}_{i=1}^{N}$
Cropping
OmiVL
$R_i = \{(t_j, r_j)\}_{j=st_i}^{ed_i}$
D

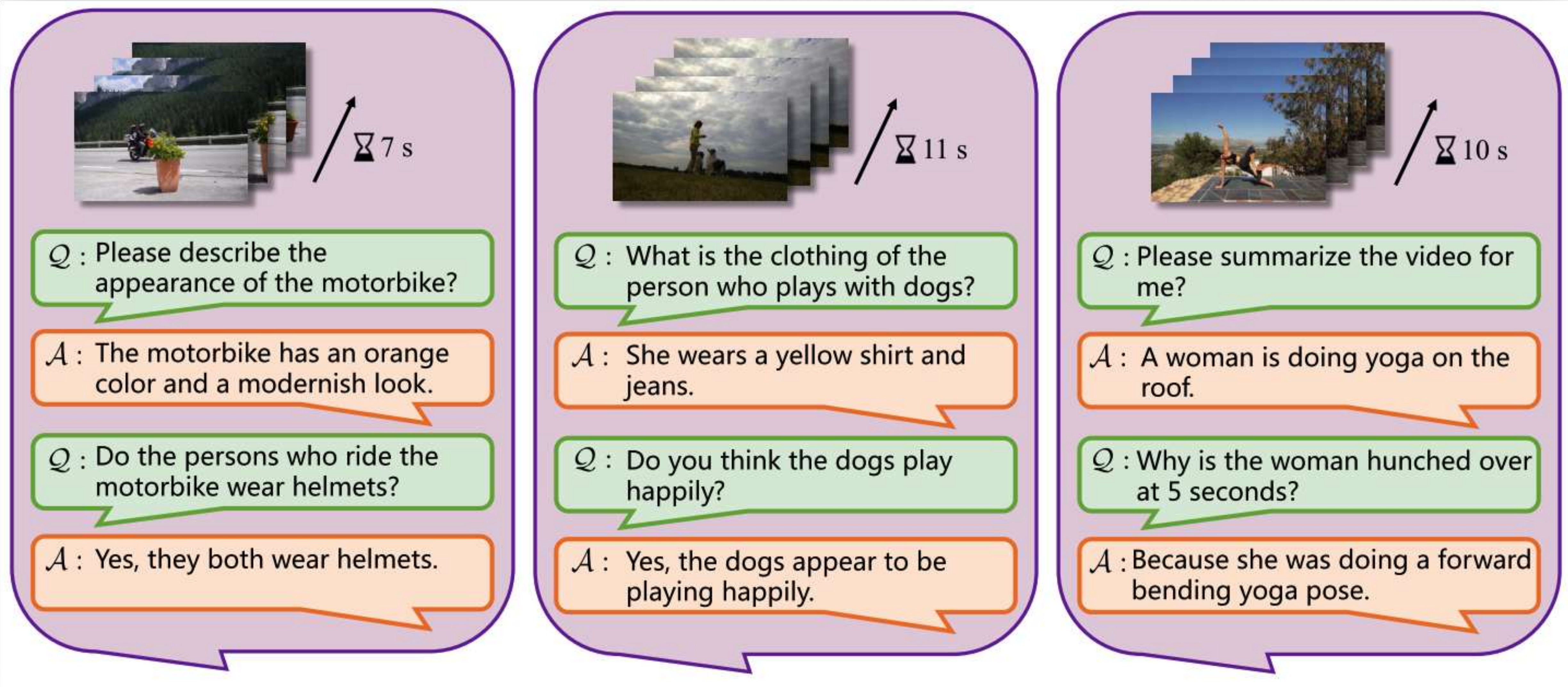


ID	Category	Appearance	Motion	Trajectory
0	environment	road and mountains	From 0 to 7 s, a motorcyclist riding on the road in the mountains; ...	N/A
1	motorcycle	orange in color	From 0 to 7 s, a man riding a motorcycle down a road; ...	at 0 s, (198,198,294,277); ...
2	person	wearing a black leather jacket and a black helmet	From 0 to 7 s, the person is a motorcyclist on a motorcycle in the mountains; ...	at 0 s, (222,176,279,259); ...
...	...	...	...	...



ID	Category	Appearance	Motion	Trajectory
0	environment	a classroom	From 0 to 1.1 s, a woman is sitting in the room; ...	N/A
1	laptop	laptop black and silver in color	From 0 to 1.1 s, a person is working on a laptop; ...	at 0 s, (181,236,289,300); ...
2	person	person long hair and green T-shirt	From 0 to 1.1 s, the person is a woman in the classroom; ...	at 0 s, (122,159,225,289); ...
3	tv	tv black screen	From 0 to 1.2 s, the tv is on a black background; ...	at 0 s, (338,133,406,181); ...
...	...	...	...	...

- ID：主键。
- Category：轨迹类别。
- Appearance：轨迹片段（实例）。
- Motion：轨迹片段（实例）的运动。
- Trajectory：轨迹片段（实例）的轨迹。
- Audio：此字段仅适用于包含完整视频的轨迹片段。



$\mathcal{Q}$: Please describe the appearance of the motorbike?

$\mathcal{A}$: The motorbike has an orange color and a modernish look.

$\mathcal{Q}$: Do the persons who ride the motorbike wear helmets?

$\mathcal{A}$: Yes, they both wear helmets.

$\mathcal{Q}$: What is the clothing of the person who plays with dogs?

$\mathcal{A}$: She wears a yellow shirt and jeans.

$\mathcal{Q}$: Do you think the dogs play happily?

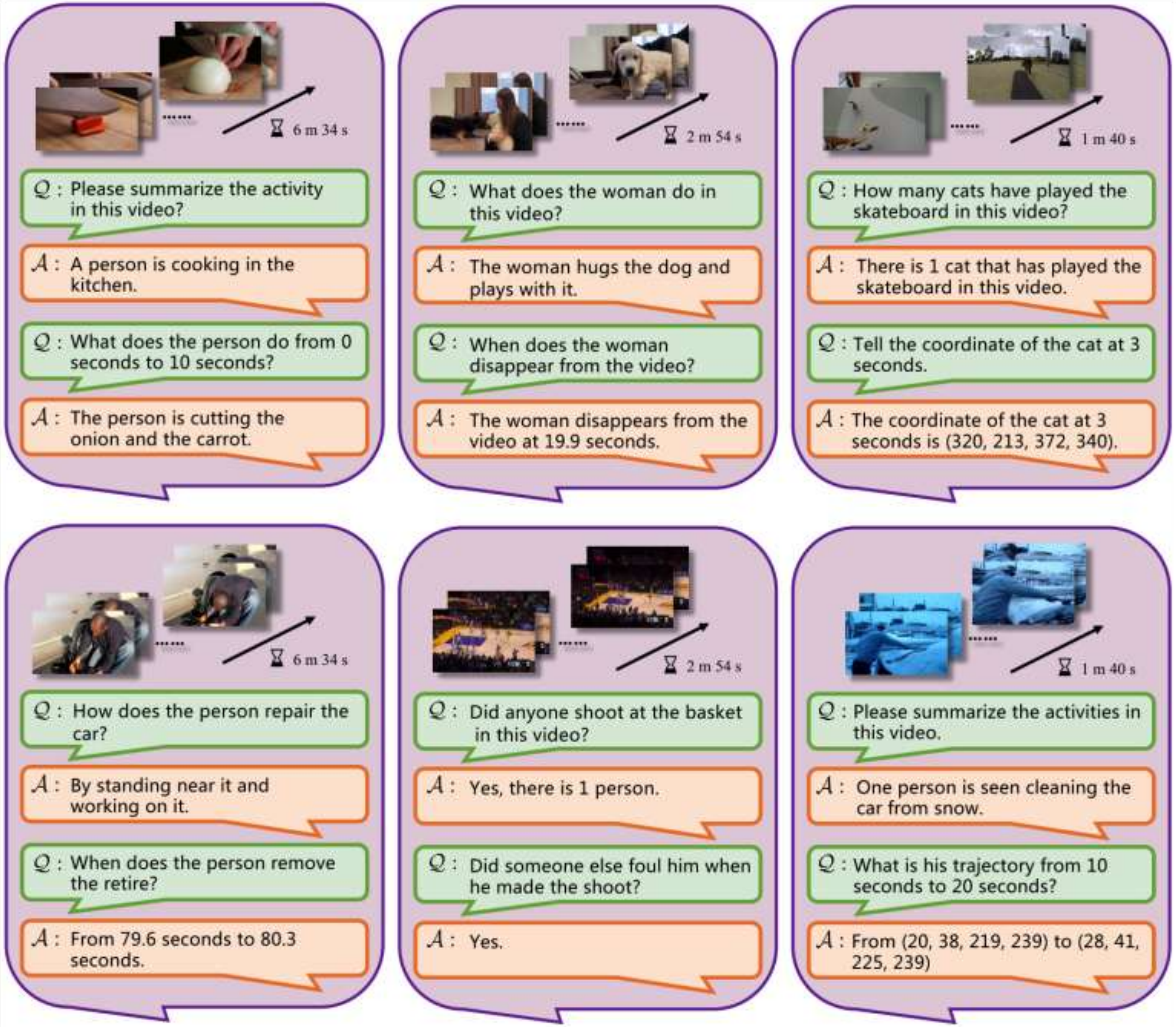
$\mathcal{A}$: Yes, the dogs appear to be playing happily.

$\mathcal{Q}$: Please summarize the video for me?

$\mathcal{A}$: A woman is doing yoga on the roof.

$\mathcal{Q}$: Why is the woman hunched over at 5 seconds?

$\mathcal{A}$: Because she was doing a forward bending yoga pose.



$\mathcal{Q}$: Please summarize the activity in this video?

$\mathcal{A}$: A person is cooking in the kitchen.

$\mathcal{Q}$: What does the person do from 0 seconds to 10 seconds?

$\mathcal{A}$: The person is cutting the onion and the carrot.

$\mathcal{Q}$: What does the woman do in this video?

$\mathcal{A}$: The woman hugs the dog and plays with it.

$\mathcal{Q}$: When does the woman disappear from the video?

$\mathcal{A}$: The woman disappears from the video at 19.9 seconds.

$\mathcal{Q}$: How many cats have played the skateboard in this video?

$\mathcal{A}$: There is 1 cat that has played the skateboard in this video.

$\mathcal{Q}$: Tell the coordinate of the cat at 3 seconds.

$\mathcal{A}$: The coordinate of the cat at 3 seconds is (320, 213, 372, 340).

$\mathcal{Q}$: How does the person repair the car?

$\mathcal{A}$: By standing near it and working on it.

$\mathcal{Q}$: When does the person remove the retire?

$\mathcal{A}$: From 79.6 seconds to 80.3 seconds.

$\mathcal{Q}$: Did anyone shoot at the basket in this video?

$\mathcal{A}$: Yes, there is 1 person.

$\mathcal{Q}$: Did someone else foul him when he made the shoot?

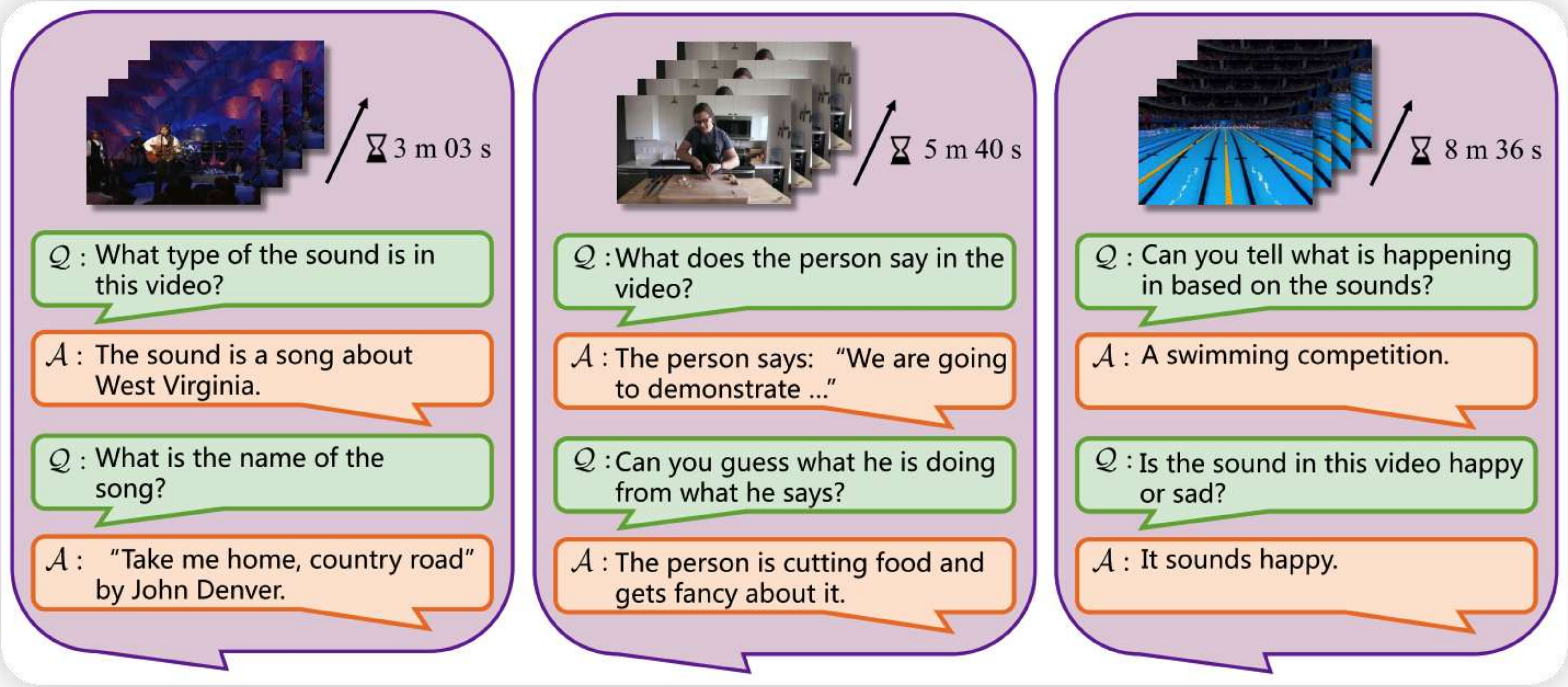
$\mathcal{A}$: Yes.

$\mathcal{Q}$: Please summarize the activities in this video.

$\mathcal{A}$: One person is seen cleaning the car from snow.

$\mathcal{Q}$: What is his trajectory from 10 seconds to 20 seconds?

$\mathcal{A}$: From (20, 38, 219, 239) to (28, 41, 225, 239)



$\mathcal{Q}$: What type of the sound is in this video?

$\mathcal{A}$: The sound is a song about West Virginia.

$\mathcal{Q}$: What is the name of the song?

$\mathcal{A}$: "Take me home, country road" by John Denver.

$\mathcal{Q}$: What does the person say in the video?

$\mathcal{A}$: The person says: "We are going to demonstrate ..."

$\mathcal{Q}$: Can you guess what he is doing from what he says?

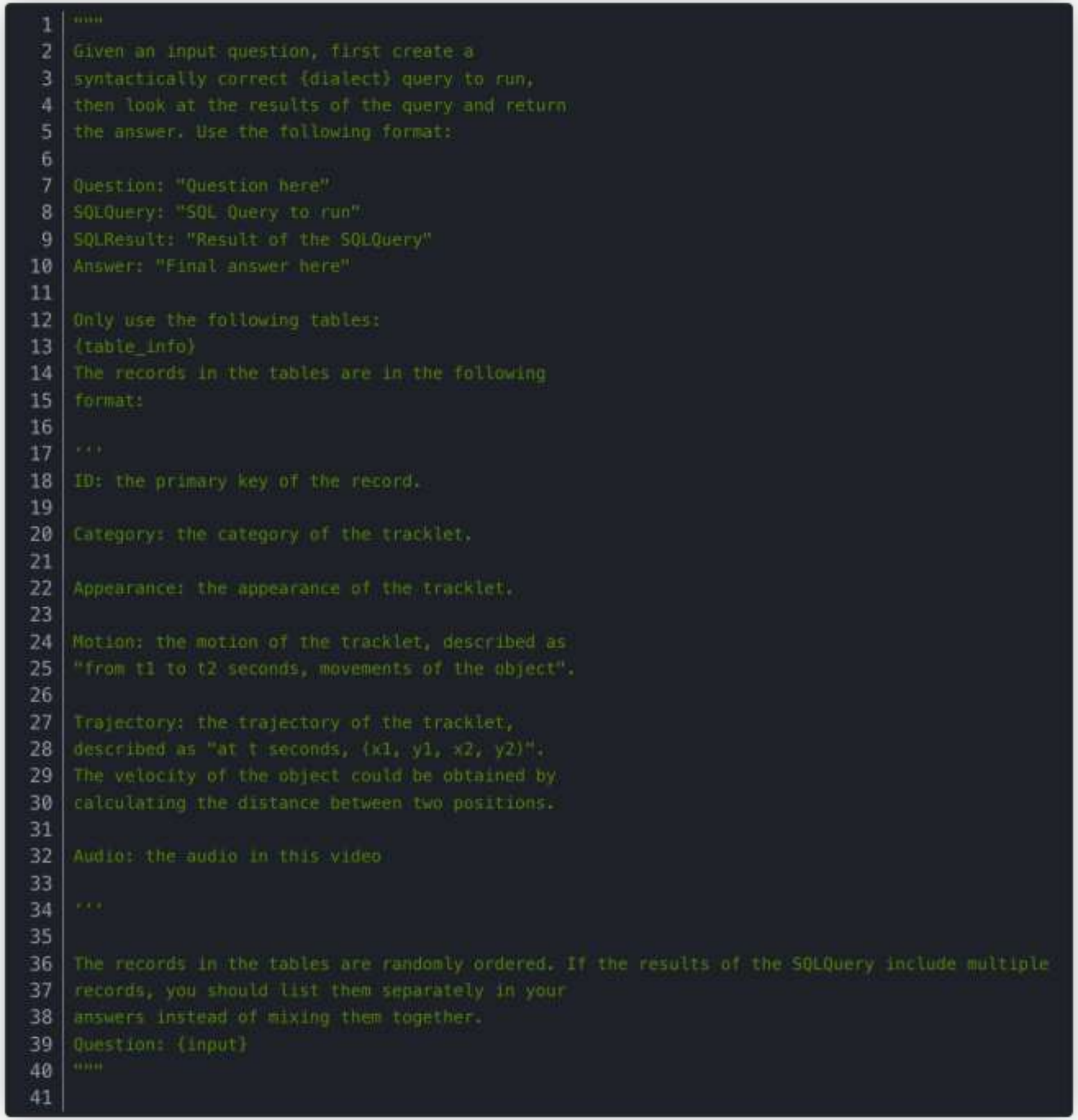
$\mathcal{A}$: The person is cutting food and gets fancy about it.

$\mathcal{Q}$: Can you tell what is happening in based on the sounds?

$\mathcal{A}$: A swimming competition.

$\mathcal{Q}$: Is the sound in this video happy or sad?

$\mathcal{A}$: It sounds happy.



```
"""
Given an input question, first create a
syntactically correct {dialect} query to run,
then look at the results of the query and return
the answer. Use the following format:

Question: "Question here"
SQLQuery: "SQL Query to run"
SQLResult: "Result of the SQLQuery"
Answer: "Final answer here"

Only use the following tables:
{table_info}
The records in the tables are in the following
format:

'''
ID: the primary key of the record.

Category: the category of the tracklet.

Appearance: the appearance of the tracklet.

Motion: the motion of the tracklet, described as
"from t1 to t2 seconds, movements of the object".

Trajectory: the trajectory of the tracklet,
described as "at t seconds, (x1, y1, x2, y2)".
The velocity of the object could be obtained by
calculating the distance between two positions.

Audio: the audio in this video

'''

The records in the tables are randomly ordered. If the results of the SQLQuery include multiple
records, you should list them separately in your
answers instead of mixing them together.
Question: {input}
"""
```

Thanks