VURF: A General-purpose Reasoning and Self-refinement Framework for Video Understanding

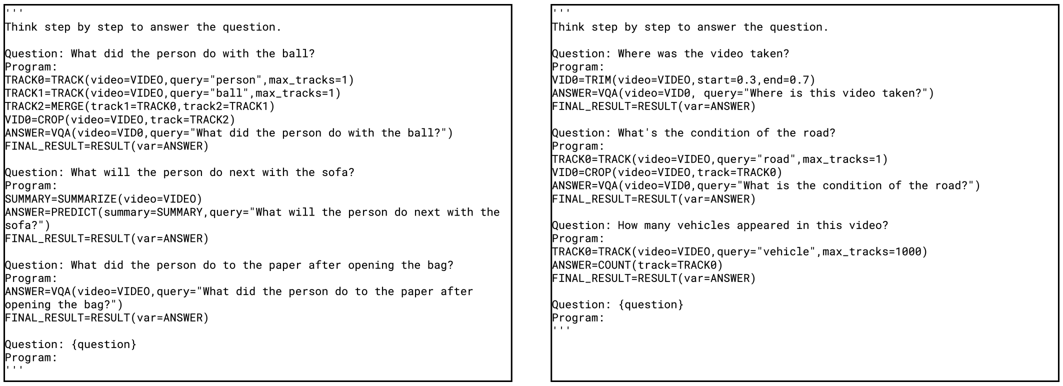
Ahmad Mahmood[1] **Ashmal Vayani**[2] **Muzammal Naseer**[3] **Salman Khan**[4,5]
amahmood@ethz.ch

Fahad Shahbaz Khan[4,6]

[1]ETH Zurich [2]University of Central Florida [3]Khalifa University, UAE
[4]Mohamed Bin Zayed University of AI, UAE [5]Australian National University, Australia
[6]Linköping University, Sweden

1 Dataset Specific Prompts

We curated in-context examples for 4 different datasets (NeXTQA, STAR, Social-IQ, and TrafficQA). The prompts for two of the datasets are shown in Figure 4. We show 3 in-context examples for each prompt.



```
'''
Think step by step to answer the question.

Question: What did the person do with the ball?
Program:
TRACK0=TRACK(video=VIDEO,query="person",max_tracks=1)
TRACK1=TRACK(video=VIDEO,query="ball",max_tracks=1)
TRACK2=MERGE(track1=TRACK0,track2=TRACK1)
VID0=CROP(video=VIDEO,track=TRACK2)
ANSWER=VQA(video=VID0,query="What did the person do with the ball?")
FINAL_RESULT=RESULT(var=ANSWER)

Question: What will the person do next with the sofa?
Program:
SUMMARY=SUMMARIZE(video=VIDEO)
ANSWER=PREDICT(summary=SUMMARY,query="What will the person do next with the
sofa?")
FINAL_RESULT=RESULT(var=ANSWER)

Question: What did the person do to the paper after opening the bag?
Program:
ANSWER=VQA(video=VIDEO,query="What did the person do to the paper after
opening the bag?")
FINAL_RESULT=RESULT(var=ANSWER)

Question: {question}
Program:
'''
```

```
'''
Think step by step to answer the question.

Question: Where was the video taken?
Program:
VID0=TRIM(video=VIDEO,start=0.3,end=0.7)
ANSWER=VQA(video=VID0, query="Where is this video taken?")
FINAL_RESULT=RESULT(var=ANSWER)

Question: What's the condition of the road?
Program:
TRACK0=TRACK(video=VIDEO,query="road",max_tracks=1)
VID0=CROP(video=VIDEO,track=TRACK0)
ANSWER=VQA(video=VID0,query="What is the condition of the road?")
FINAL_RESULT=RESULT(var=ANSWER)

Question: How many vehicles appeared in this video?
Program:
TRACK0=TRACK(video=VIDEO,query="vehicle",max_tracks=1000)
ANSWER=COUNT(track=TRACK0)
FINAL_RESULT=RESULT(var=ANSWER)

Question: {question}
Program:
'''
```

Figure 1: **LEFT:** LLM prompt for *STAR* dataset **RIGHT:** LLM prompt for *TrafficQA* dataset

2 Refinement Prompts

2.1 Error Correction

The error correction module queries the LLM two times, one to receive feedback on a program and the other to correct the program given the feedback. The prompts are shown in Figure 2. We display 2 in-context examples for each prompt.

2.2 Self-Refinement

The self-refinement module (which we pre-apply to refine our in-context examples) consists of 2 major queries to the LLM. One is for generating a context-free program to avoid hallucinations. Other is to convert the context-free program to a valid program. The prompts are shown in Figure 3

Workshop on Video-Language Models at 38th Conference on Neural Information Processing Systems (NeurIPS 2024).

```
'''
Give feedback on the provided instruction/program pair. The list of valid
available functions is [VQA, TRACK, SEGMENT, GROUND, SUMMARISE, POSE_DETECT,
MERGE, CROP, TRIM, TRIMAFTER, TRIMBEFORE, PREDICT, COUNT,
EVAL,COLORPOP,BGBLUR]

Instruction: What did the man do after sitting down?
Program:
INTERVAL0=LOCALISE(video=VIDEO,query="man sits down")
VID0=TRIMAFTER(video=VIDEO,interval=INTERVAL0)
ANSWER=VQA(video=VID0,query="What is the man doing?")
Feedback:
The function LOCALISE is not available. So the program is incorrect.

Instruction: Make a colorpop of the man who is running
Program:
TRACK0=TRACK(video=VIDEO,query="running man")
RESULT=COLORPOP(video=VIDEO,track=TRACK0)
Feedback:
The program is correct.

Instruction: {instruction}
Program:
{program}
Feedback:
'''
```

```
'''
Given the feedback, give the correct program. The list of valid available
functions is [VQA, TRACK, SEGMENT, GROUND, SUMMARISE, POSE_DETECT, MERGE,
CROP, TRIM, TRIMAFTER, TRIMBEFORE, PREDICT, COUNT, EVAL,COLORPOP,BGBLUR]

Instruction: What did the man do after sitting down?
Program:
INTERVAL0=LOCALISE(video=VIDEO,query="man sits down")
VID0=TRIMAFTER(video=VIDEO,interval=INTERVAL0)
ANSWER=VQA(video=VID0,query="What is the man doing?")
Feedback:
The function LOCALISE is not available. So the program is incorrect.
Correct Program:
INTERVAL0=GROUND(video=VIDEO,query="man sits down")
VID0=TRIMAFTER(video=VIDEO,interval=INTERVAL0)
ANSWER=VQA(video=VID0,query="What is the man doing?")

Instruction: Make a colorpop of the man who is running
Program:
TRACK0=TRACK(video=VIDEO,query="running man")
RESULT=COLORPOP(video=VIDEO,track=TRACK0)
Feedback:
The program is correct.
Correct Program:
TRACK0=TRACK(video=VIDEO,query="running man")
RESULT=COLORPOP(video=VIDEO,track=TRACK0)

Instruction: {instruction}
Program:
{program}
Feedback:
{feedback}
Correct Program:
'''
```

Figure 2: **LEFT:** The prompt for the feedback generation of a given program. **RIGHT:** Given a feedback the correct program is generated by giving this prompt to the LLM.

```
'''
Think step by step to answer the question. Use any functions but make sure to keep the structure of the program like the following example:

Question: Where was the video taken?
Program:
VID0=TRIM(video=VIDEO,start=0.3,end=0.7)
ANSWER=VQA(video=VID0, query="Where is this video taken?")
FINAL_RESULT=RESULT(var=ANSWER)

Question: {question}
Program:

'''
```

```
'''
Given a perfect program and an old program, give a new program that uses valid functions. The list of valid available functions is [VQA, TRACK, SEGMENT,
GROUND, SUMMARISE, POSE_DETECT, MERGE, CROP, TRIM, TRIMAFTER, TRIMBEFORE, PREDICT, COUNT, EVAL,COLORPOP,BGBLUR]

Examples of Valid Programs:
SUMMARY=SUMMARIZE(video=VID0)
ANSWER=PREDICT(summary=SUMMARY,query="What will the person do next with the ball?")
FINAL_RESULT=RESULT(var=ANSWER)

VID0=TRIM(video=VIDEO, interval=(0.5,1))
TRACK0=TRACK(video=VID0, query="Man in black jacket moving his legs", max_tracks=1)
VID1=CROP(video=VID0, track=TRACK0)
ANSWER=VQA(video=VID1, query="How is the man moving his hands?")

INTERVAL=GROUND(video=VIDEO, query="Baby puts down the ipod")
VID0=TRIM_AFTER(video=VIDEO, interval=INTERVAL)
ANSWER=VQA(video=VID0, query="Why does the baby pick up the ipod?")

Old Program:
{old_program}
Perfect Program:
{perfect_program}
Program:
'''
```

Figure 3: **TOP:** The prompt for the context-free generation. **BOTTOM:** The prompt for aligning a "perfect" program to a valid program

3 Ablations

3.1 LLM

We conducted experiments on two datasets (STAR and Social-IQ-2.0) using three additional LLM models for program generation: GPT-4o, CodeLLAMA, and Code-T5. The results, shown in Table 1, indicate that Instruct models excel in program generation due to the task's instruction-oriented nature. Specifically, CodeLLAMA-7b performed poorly, generating numerous errors likely because it is trained to produce actual code rather than pseudocode-like instructions. Code-T5 failed to generate a correct program altogether, resulting in minimal performance.

3.2 Transcripts

Since the Social-IQ-2.0 dataset contains the transcripts of the videos so we introduce a new function, ANALYSE, tailored for questions aimed at discerning the mood or tone of the video conversation. This function involves querying a Language Model (LLM) with the transcript, and, in some instances,

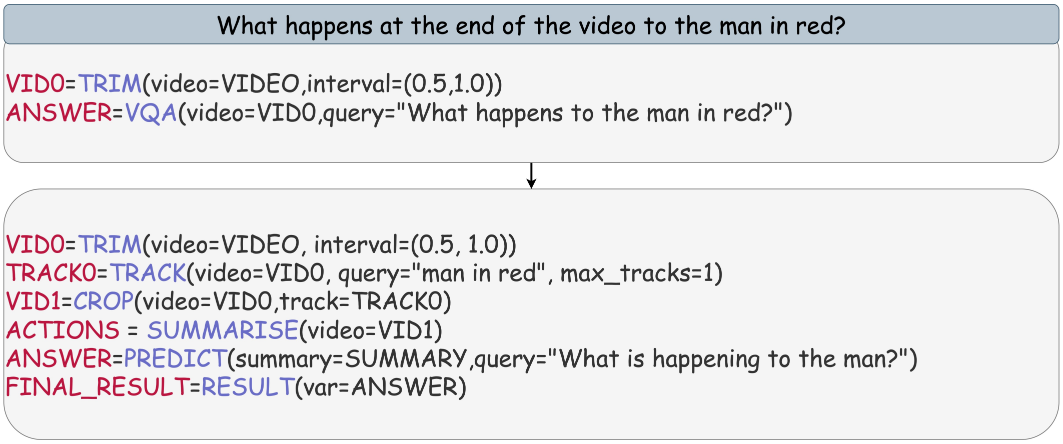
Table 1: ***Comparison with multiple LLMs***, both open source models and other GPT model versions.

LLM	STAR	Social-IQ-2.0
GPT-3.5-turbo-instruct (original)	47.2%	51.6%
GPT-4	43.5%	49.1%
CodeLlama-7b	10.0%	5.0%
CodeLlama-7b-Instruct	42.3%	48.4%
Code-T5	0%	0%

Table 2: ***Effect of using transcripts from the Social-IQ QA dataset***.

Dataset	Baseline	With transcripts
Social-IQ-2.0	48.1%	51.6%

the appended summary along with the posed question. Our analysis includes performance metrics for both scenarios i.e., employing transcripts and not using them (Table 2).

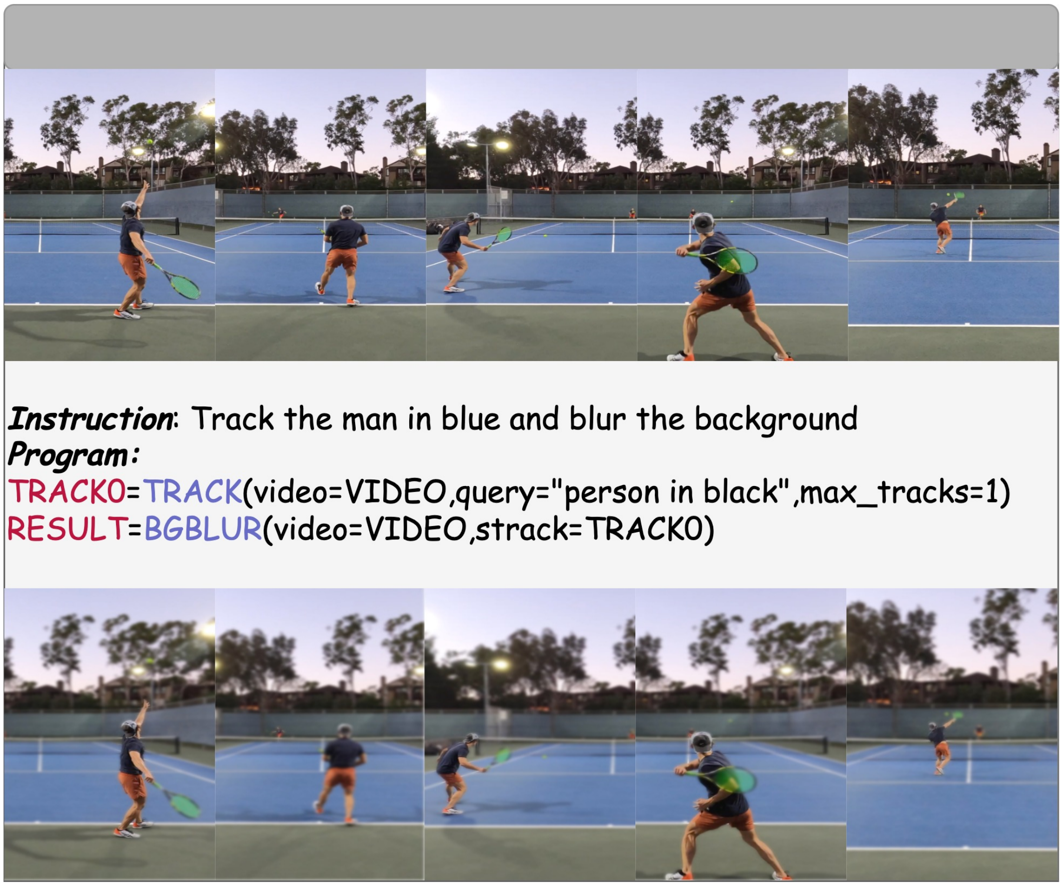


What happens at the end of the video to the man in red?

```
VID0=TRIM(video=VIDEO,interval=(0.5,1.0))
ANSWER=VQA(video=VID0,query="What happens to the man in red?")
```

```
VID0=TRIM(video=VIDEO, interval=(0.5, 1.0))
TRACK0=TRACK(video=VID0, query="man in red", max_tracks=1)
VID1=CROP(video=VID0,track=TRACK0)
ACTIONS = SUMMARISE(video=VID1)
ANSWER=PREDICT(summary=SUMMARY,query="What is happening to the man?")
FINAL_RESULT=RESULT(var=ANSWER)
```

Figure 4: An example of the self refinement module. An initially generated prompt is refined by the LLM to produce a more complex and modular program.

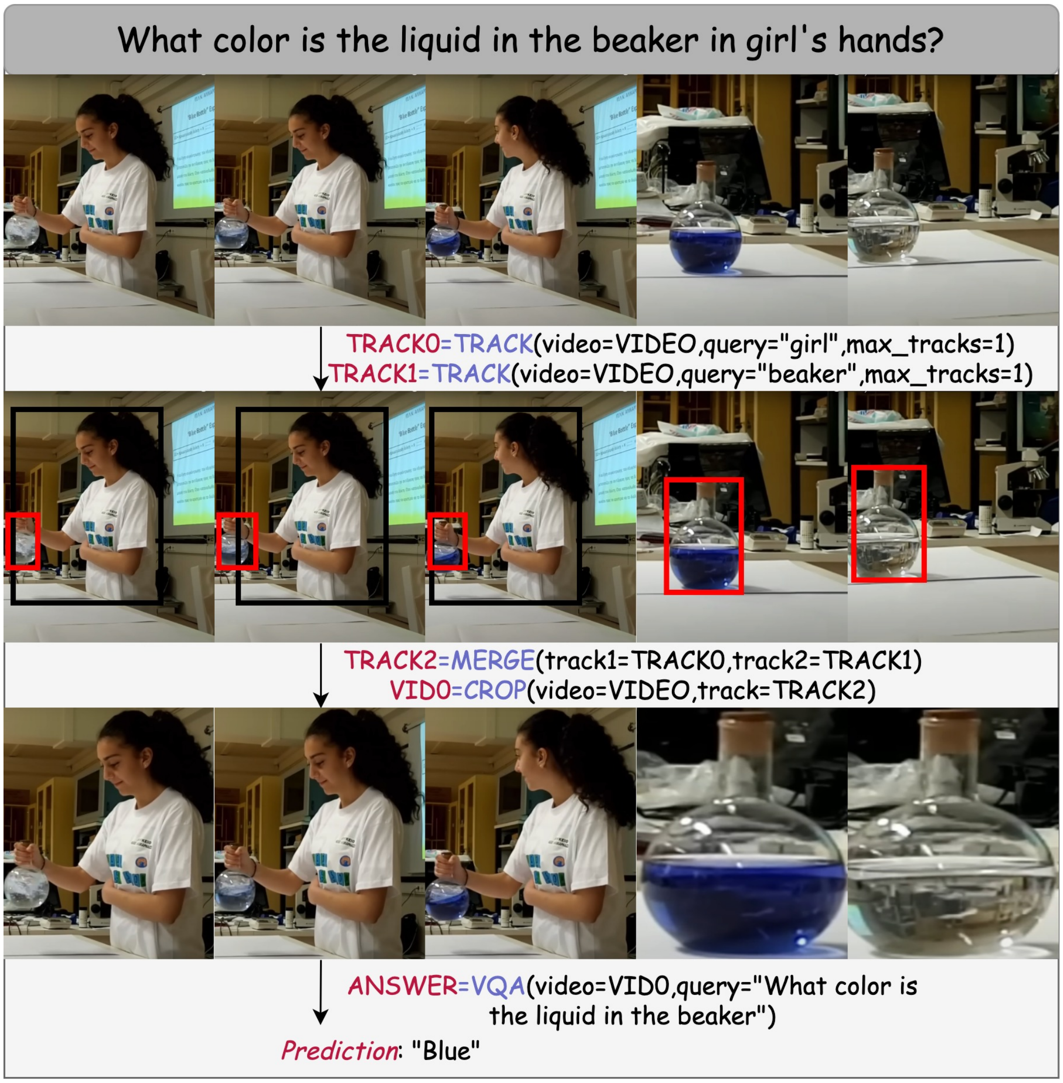


Instruction: Track the man in blue and blur the background

Program:

```
TRACK0=TRACK(video=VIDEO,query="person in black",max_tracks=1)
RESULT=BGBLUR(video=VIDEO,strack=TRACK0)
```

Figure 5: A qualitative example of video-editing using VURF.



What color is the liquid in the beaker in girl's hands?

```
TRACK0=TRACK(video=VIDEO,query="girl",max_tracks=1)
TRACK1=TRACK(video=VIDEO,query="beaker",max_tracks=1)
```

```
TRACK2=MERGE(track1=TRACK0,track2=TRACK1)
VID0=CROP(video=VIDEO,track=TRACK2)
```

```
ANSWER=VQA(video=VID0,query="What color is
                the liquid in the beaker")
```

Prediction: "Blue"

Figure 6: A step by step qualitative example of Video Question Answering using VURF.