Vript: A Video Is Worth Thousands of Words

Dongjie Yang[1], Suyuan Huang[2], Chengqiang Lu[3], Xiaodong Han[3], Haoxin Zhang[3], Yan Gao[3], Yao Hu[3], Hai Zhao[1,*]
[1]Shanghai Jiao Tong University, [2] Beihang University, [3] Xiaohongshu Inc.

Neurips 2024

Key Contributions

1. Large-scale Video-text Dataset

We open-source three large-scale video-text dataset, including **Vript**, **Vript_CN** and **Vript_Multilingual**. These three dataset cover different UGC video sources, containing high-quality videos and corresponding captions.

2. New paradigms of video-text alignment

We explore three paradigms of video-text alignment in large multimodal models based on Vript dataset, building the powerful video captioning model with SOTA performance.

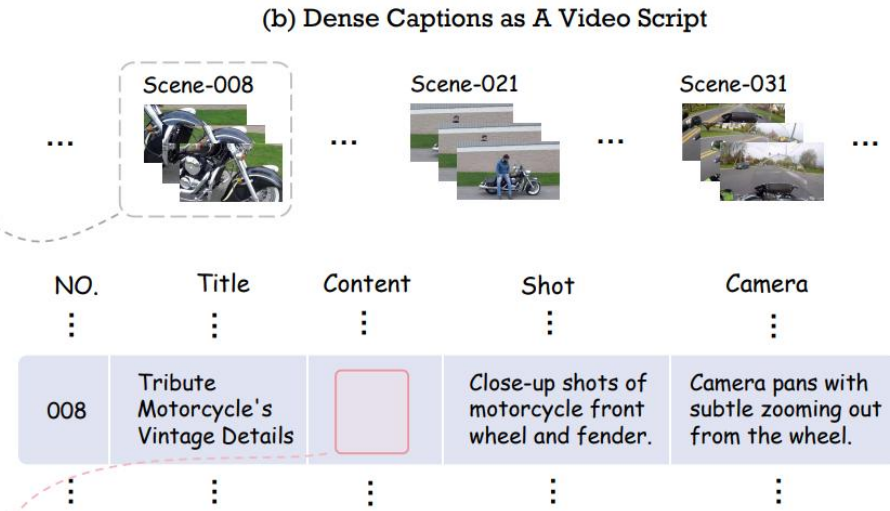
3. A challenging video understanding benchmark.

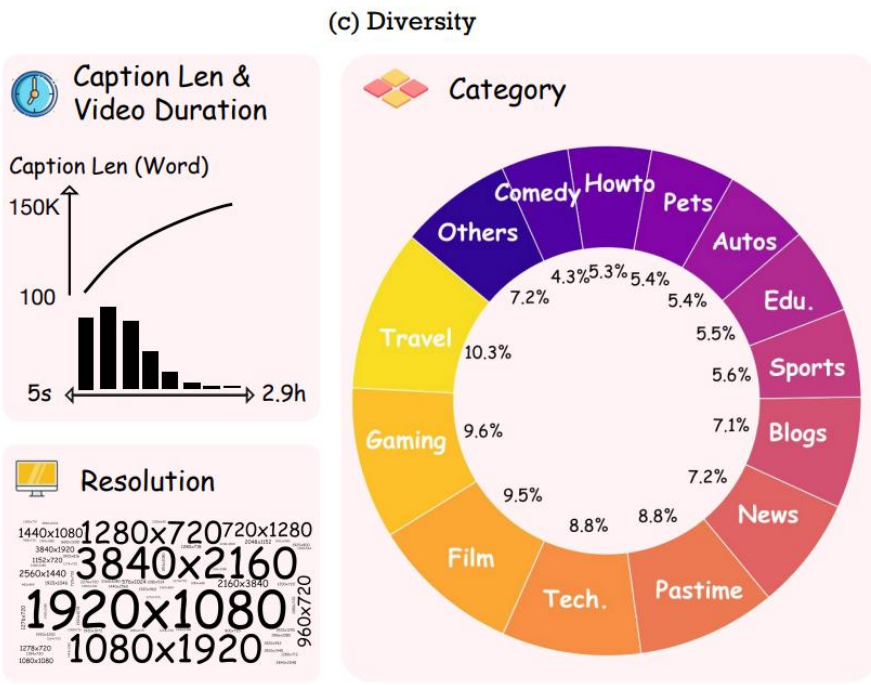
We construct three different challenging benchmarks for evaluating next-level video LLMs.

Vript dataset

- Very detailed (over 150 words)
- Consistent clips (long video have multiple clips)
- high resolution (720p-2K)
- diverse categories

Vript
A large-scale video-text dataset of high-resolution videos ann
Mutonix/Vript
Viewer • Updated Jun 11 • 409k • 6.1k • 14





Comparsion with other datasets

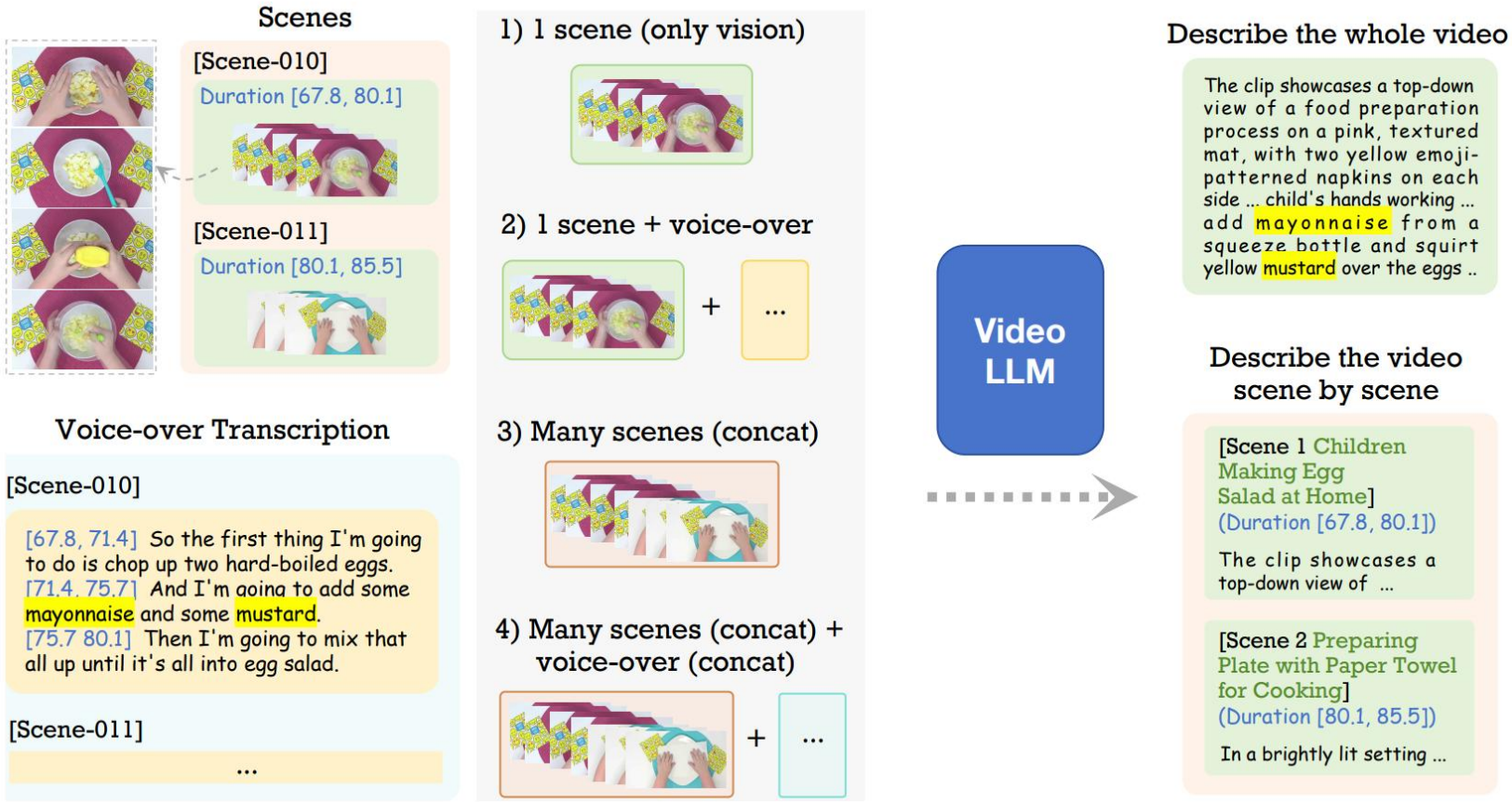
Dataset	Domain	Text Len	Clips	Duration	Resolution	Lang
HowTo100M [21]	Open	4.0	136M	134Kh	240p	en
ACAV100M [22]	Open	-	100M	278h	-	en
HD-VILA-100M [23]	Open	32.5	103M	371Kh	720p	en
WebVid-10M [13]	Open	~12	10M	~52Kh	360p	en
YT-Temporal-180 [24]	Open	~10	180M	-	480p	en
MSVD [25]	Open	8.7	1970	5.3h	-	en
MSR-VTT [16]	Open	9.3	10K	40h	240p	en
DiDeMo [26]	Flickr	8.0	27K	87h	-	en
ActivityNet [27]	Action	13.5	100K	849h	144p-720p	en
YouCook2 [28]	Cooking	8.8	14K	176h	-	en
VATEX [29]	Open	15.2	41K	~115h	-	en
HD-VG-130M [6]	Open	~10	130M	~180Kh	720p	en
Panda-70M [14]	Open	13.2	70M	167Kh	720p	en
InternVid [30]	Open	17.6	234M	760.3Kh	720p	en
Vript	Open	~145	420K	1.3Kh	720p-2K	en
Vript-CN	Open	~150	293K	-	720p-1080p	zh
Vript-Multilingual	Open	~150	677K	-	720p-1080p	multi

Video-text Alignment

1. Video-Script alignment

2.Voice-over Transcription

3. Video Timestamp



Capability of Vriptor-STLLM

Table 2: Different strategies of video-script alignment and voice-over transcription.

Strategy	Vript-HAL			MSR-VTT
	Precision	Recall	F1	Recall
2 scenes	**75.8**	40.9	53.1	122.0
3 scenes	74.1	49.5	59.4	135.8
4 scenes	72.3	55.8	63.0	138.1
5 scenes	71.4	**57.5**	**63.7**	**139.5**
Whole	79.1	26.8	40.0	83.0
Whole (voice)	**80.3**	**27.7**	**41.1**	-

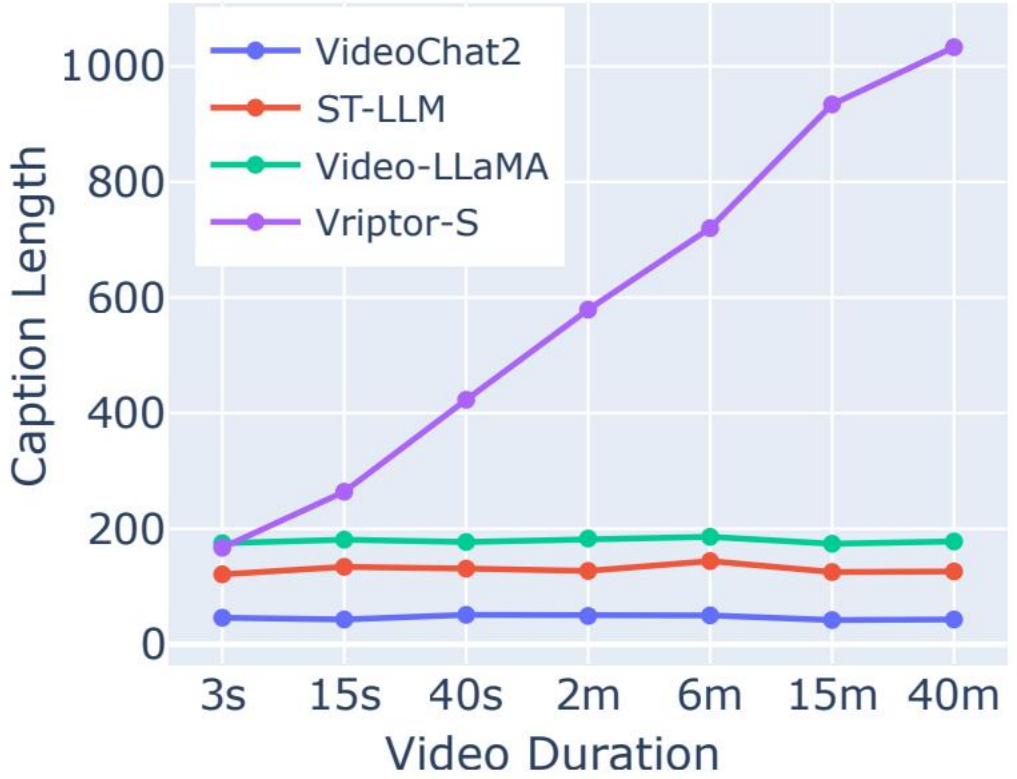


Figure 3: Caption lengths for videos of different durations.

Vript-HAL: first benchmark evaluating action and object hallucinations in video LLMs

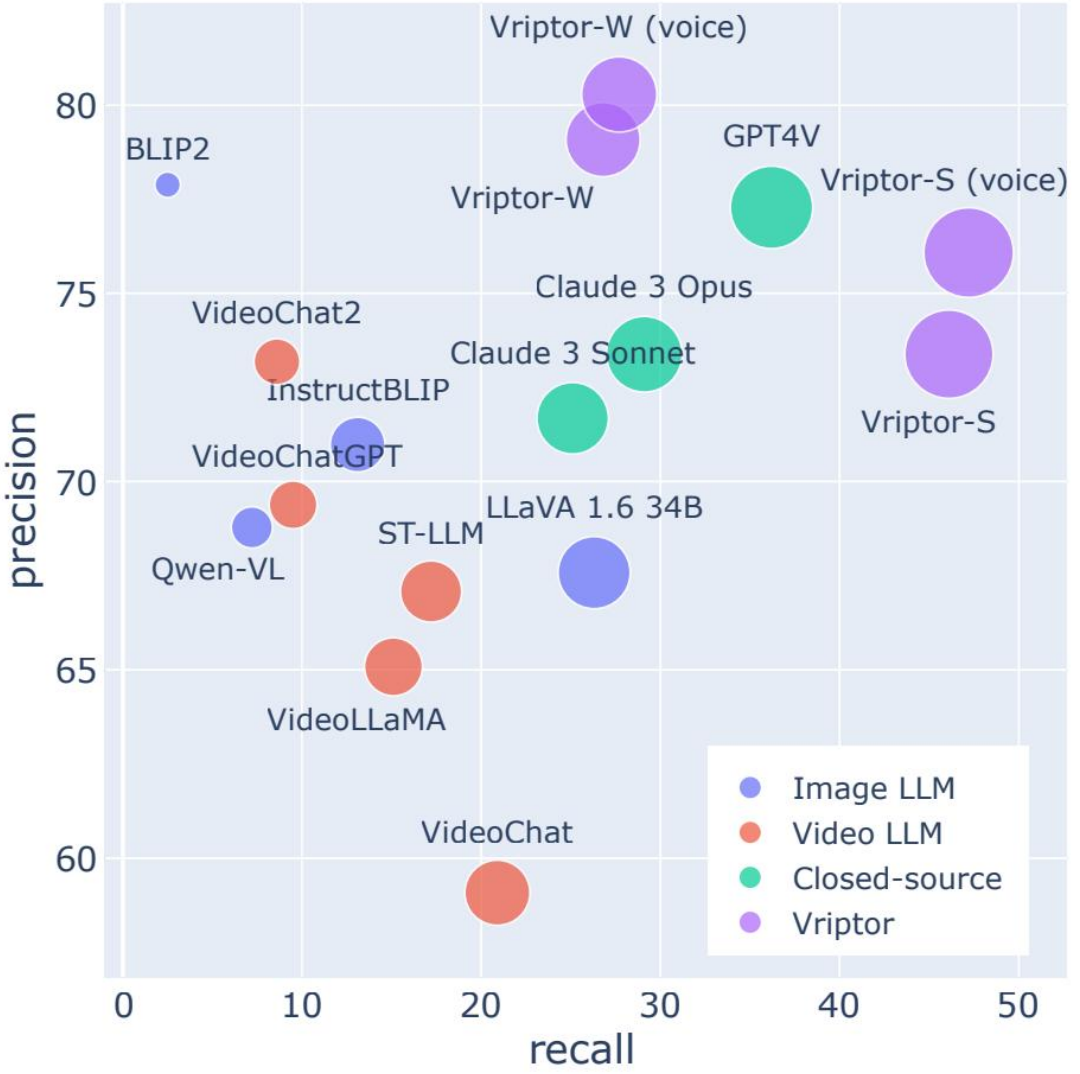
Panda-70M

The view from the handlebars of a motorcycle as it drives down a street.

Vript-HAL

The video clip features a rider operating a motorcycle, presumably an Indian brand bike, from a first-person perspective. The bike has a classic design with chrome detailing and black leather elements. The rider's hands are visible, wearing bright gloves, manipulating the bike's controls. We see a clear windshield, a well-kept dashboard with gauges, and the front part of the bike, including the headlight, which may have extra lights - a detail the voice-over is uncertain about. The environment is a suburban street during the daytime with green lawns, houses, and passing cars. The sky is overcast. The rider comments on starting the bike with the choke and their perplexity about the 'Indian' aspect of the motorcycle, while also noting its coolness. There are no other characters in sight; the focus is on the rider's experience and interaction with the bike.

$$\mathcal{P}(\mathbf{p},\mathbf{g}) = \frac{\#\{\mathbf{p}\cap\mathbf{g}\}}{\#\{\mathbf{p}\}}, \quad \mathcal{R}(\mathbf{p},\mathbf{g}) = \frac{\#\{\mathbf{p}\cap\mathbf{g}\}}{\#\{\mathbf{g}\}}, \quad F_1 = 2\cdot\frac{\mathcal{P}\cdot\mathcal{R}}{\mathcal{P}+\mathcal{R}},$$



Vript-RR: combining reasoning with retrieval resolving question ambiguity in long-video QAs

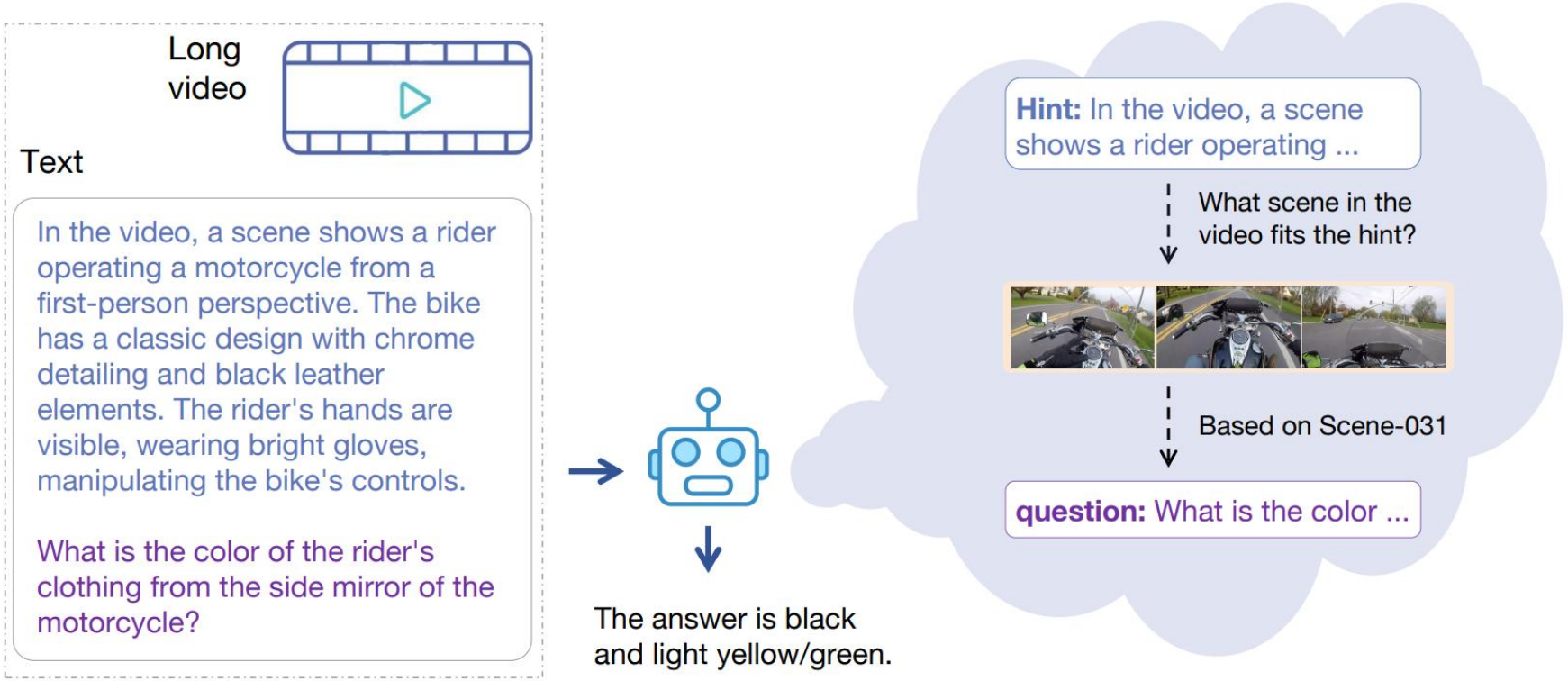


Figure 5: The overview of answering the question in Vript-RR, which is an end-to-end process.

Vript-ERO: evaluate the temporal understanding of events in long videos rather than actions in short videos

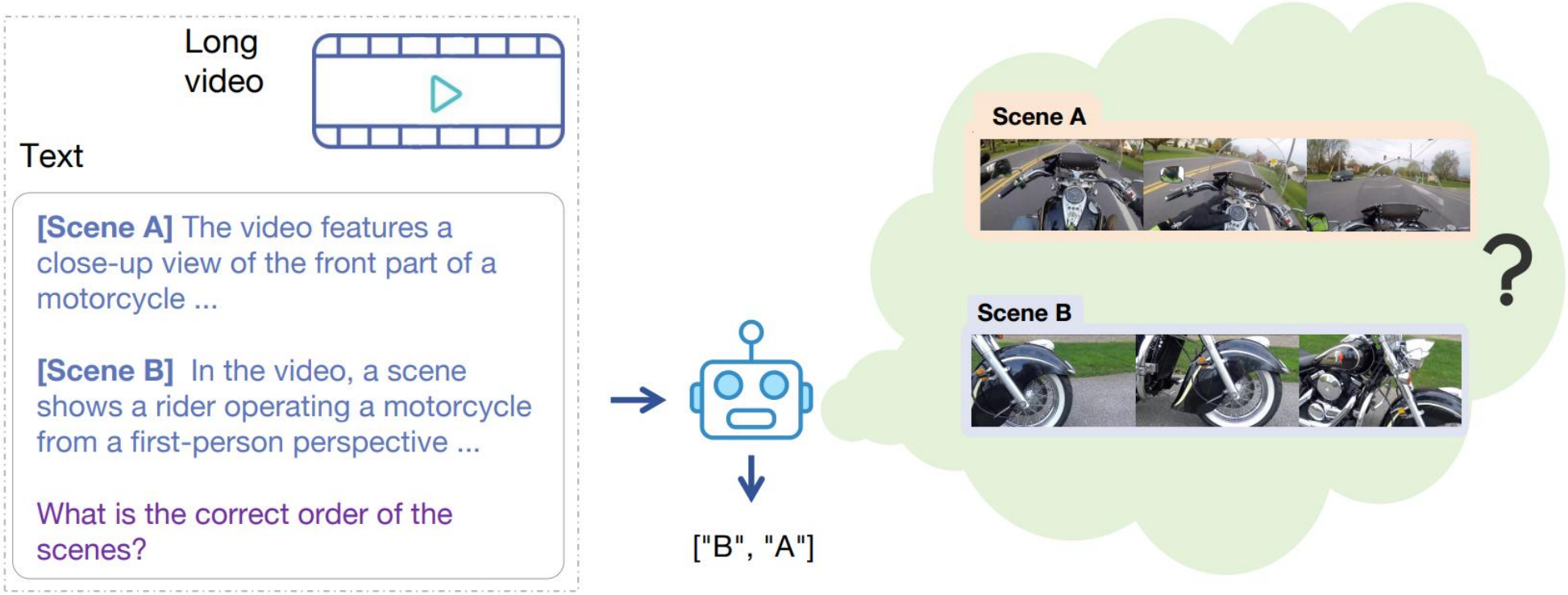
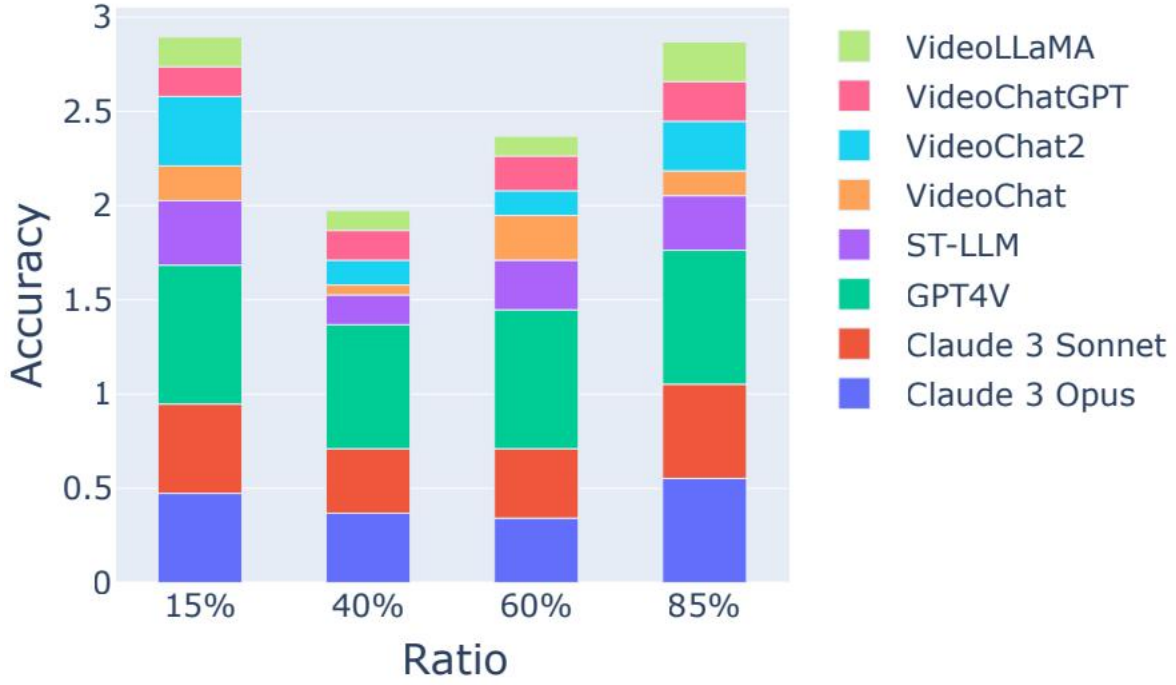
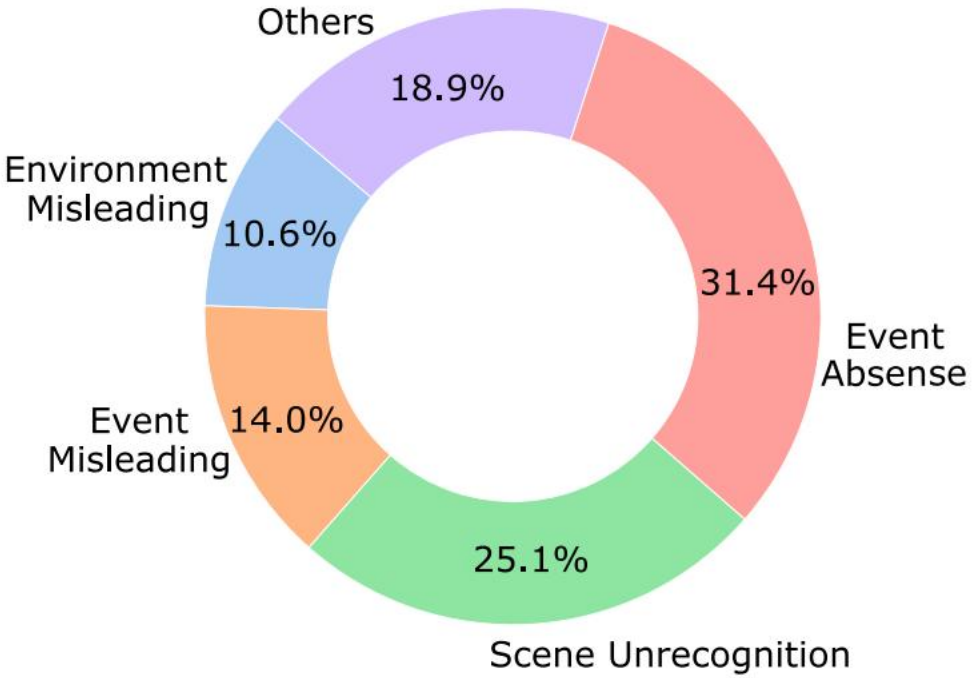


Figure 7: The overview of answering the question in Vript-ERO.

Needle-in-a-TimeStack & Failure in Vript-ERO



(a)



(b)