Flash-VStream: Efficient Real-Time Understanding for Long Video Streams

Supplementary Material

In the supplementary material, we first provide implementation details of the Flash Memory mechanism and training settings. Subsequently, we conduct an analysis experiment on model inference efficiency and more ablation studies on memory structure configurations. We then present more visual cases to provide a comprehensive understanding of the performance of models. We highly recommend watching **the supplementary video**, which contains a live demonstration of real-time multimodal assistant based on Flash-VStream model.

A. Implementation Details

This section describes the details of the proposed Flash Memory mechanism in Sec. 3. The Flash Memory consists of Context Synopsis Memory (CSM) and Detail Augmentation Memory (DAM). CSM uses a clustering-based updating policy, while DAM uses a retrieval-based updating policy.

$$M_k^{\text{CSM}} = \frac{1}{|S_k|}\sum_{i\in S_k} e_i^{\text{L}}, 1 \le k \le N^{\text{CSM}} \tag{1}$$

$$M^{\text{CSM}} = \text{cluster}(M^{\text{CSM}} \oplus e_{t+1}^{\text{L}}) \tag{2}$$

A.1. Context Synopsis Memory

As mentioned in Sec. 3.2, CSM is designed for aggregating long-context temporal information and modeling the distribution of information density. M_k^{CSM} represents the centroid of the k-th cluster. M^{CSM} is initialized with the first N^{CSM} feature maps of the first N^{CSM} frames. When the next frame arrives, a clustering algorithm is employed to consolidate its feature map into existing clusters. Here we illustrate the "cluster" operation of Eq. (2) in detail.

As shown in Alg. 1, CSM performs a temporal-wise *K-means Clustering* algorithm to condense $(N^{\text{CSM}}+1) \times h' \times w'$ tokens to $N^{\text{CSM}} \times h' \times w'$ tokens. Each frame feature in temporal memory $M_k^{\text{CSM}} = c_k \in \mathbb{R}^{h' \times w' \times d}$ represents the centroid of the i-th feature map cluster.

A.2. Detail Augmentation Memory

As described in Sec. 3.3, DAM aims at storing spatial details of the most informative key frames, based on the feature clusters of CSM. For DAM, we use a Feature-Centric Sampling method to calculate $M^{\text{DAM}} \in \mathbb{R}^{N^{\text{DAM}} \times h \times w \times d}$.

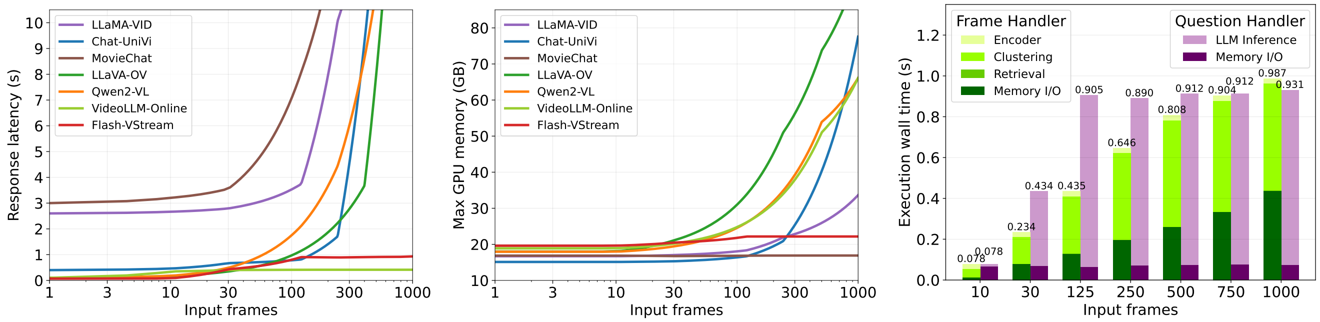
Alg. 2 shows the pseudo code of *Feature-Centric Retrieval*. Here w_j is equal to the size of j-th cluster, i.e., the number of feature maps in this cluster. We choose the centroids of the top-k largest clusters as anchors. Then we select key features from the feature bank E_t^H. E_t^H keeps high-resolution feature maps of all frames on disk, where t is the current number of frames. The features nearest to these anchors in the feature map space are considered as key features, which are added to the DAM.

Algorithm 1 K-means Clustering Algorithm

```
Require: Current cluster centroids $M = M^{\text{CSM}}$
Require: Newest frame feature $e = e_t^{\text{L}}$
Require: Set of all points $X = \{M_1, M_2, \ldots, M_N, e\}$
Require: Weights vector of points $W = \{w_1, w_2, \ldots, w_N, 1\}$
Require: Maximum memory length $N = N^{\text{CSM}}$
Require: Maximum number of iterations $T$
procedure K-MEANS($X, W, N, T$)
    Initialize $t \leftarrow 0$
    Initialize centroids $C = \{c_1, c_2, \ldots, c_N\}$ from $X$
    Initialize cluster assignment $S_j \leftarrow \{\}, 1 \le j \le N$
    while $t < T$ do
        for $x_i \in X$ do
            $j \leftarrow \underset{j}{\text{argmin}} \|x_i - c_j\|^2$
            $S_j \leftarrow S_j \cup \{x_i\}$
        end for
        for $j = 1, 2, \ldots, N$ do
            $c_j^{\text{new}} \leftarrow \frac{\sum_{x_i \in S_j} w_i \cdot x_i}{\sum_{x_i \in S_j} w_i}$
        end for
        Clear $S$
        $C \leftarrow C^{\text{new}}$
        $t \leftarrow t + 1$
    end while
    for $j = 1, 2, \ldots, N$ do
        $w_j^{\text{CSM}} \leftarrow \sum_{x_i \in S_j} w_i$
    end for
    $M^{\text{CSM}} = C$
    $W^{\text{CSM}} = \{w_1^{\text{CSM}}, w_2^{\text{CSM}}, \ldots, w_N^{\text{CSM}}\}$
    return $M^{\text{CSM}}, W^{\text{CSM}}$
end procedure
```

Algorithm 2 Feature-Centric Sampling

```
Require: Current feature bank $E_t^{\text{H}} = \{e_1^{\text{H}}, e_2^{\text{H}}, \ldots, e_t^{\text{H}}\}$
Require: Current cluster centroids $M^{\text{CSM}}$
Require: Weights vector of points $W = \{w_1, w_2, \ldots, w_N\}$
Require: Maximum memory length $N = N^{\text{DAM}}$
procedure KEY FEATURE RETRIEVAL($E^{\text{H}}, M^{\text{CSM}}, W, N$)
    $k \leftarrow N$
    $idx \leftarrow$ argsort($W$, descending=True)
    $j_1, j_2, \ldots, j_k \leftarrow idx[: k]$
    $M^{\text{DAM}} \leftarrow \{\}$
    for $z = 1, 2, \ldots, k$ do
        $anchor \leftarrow M_{j_z}^{\text{CSM}}$
        $i \leftarrow \underset{i \le t}{\text{argmin}} \|e_i^{\text{L}} - anchor\|^2$
        $M^{\text{DAM}} \leftarrow M^{\text{DAM}} \cup \{e_i^{\text{H}}\}$
    end for
    return $M^{\text{DAM}}$
end procedure
```



(a) Response latency (s) v.s. Input frames

(b) Max GPU memory (GB) v.s. Input frames

(c) Execution wall time (s) v.s. Input frames

Figure 1. **(a) Response latency comparison. (b) Max GPU memory comparison. (c) Execution wall time analysis.** Response latency refers to the wall time between inputting a question and outputting the first token of the answer. Max GPU memory indicates the peak GPU memory usage during inference. All experiments were conducted on A100 GPUs using BFloat16 and FlashAttention-2.

Settings	Value
Batch Size	64
Learning Rate	8e-4
Lora Rank	64
Lora Alpha	32
Learning Schedule	Cosine decay
Warmup Ratio	0.01
Weight Decay	0.1
Epoch	1
Optimizer	AdamW
Deepspeed Stage	2
Visual Encoder	Freeze
Projector	Open
LLM	Open

Table 1. Training settings of Flash-VStream.

B. Training Details

We train Flash-VStream on a 9k subset of LLaVA-Video [11] dataset for one epoch. During training, we freeze the parameters of visual encoder, while all linear layers of projector and LLM are LoRA finetuned. The overall training can be finished in about 10 hours on 8 A100 80G GPUs with BFloat16 automatic mixed precision and FlashAttention-2 [1]. Detailed training settings are shown in Tab. 1.

C. Efficiency Analysis

An efficiency analysis is performed to assess the inference efficiency of Flash-VStream. Specifically, we concentrate on the response latency and GPU memory consumption of models, as discussed in Sec. 1 of the paper.

We compare Flash-VStream with other competitive video language models [2, 4, 5, 7, 8] in terms of response latency and max GPU memory. As presented in Fig. 1, Flash-VStream demonstrates superior performance in both efficiency metrics. Fig. 1a shows the response latency comparison, where Flash-VStream consistently exhibits lower latency across varying numbers of input frames. This indicates that Flash-VStream is more efficient in processing video inputs, resulting in faster response times (less than 1 second). Fig. 1b illustrates the maximum GPU memory usage. Flash-VStream maintains a relatively stable and lower GPU memory consumption compared to other models, even as the number of input frames increases. This efficiency in memory usage makes Flash-VStream more scalable and suitable for deployment in resource-constrained environments.

From a systematic perspective, we measure the execution wall time of each process in Fig. 1c. The result shows that the question handler process stays fast enough ($<$ 1s) regardless of the number of input frames. This is because the question handler only relies on size-fixed Flash Memory. The execution time of the frame handler process grows up to more than 1 second when the number of frames exceeds 1000. Although this may result in delayed updates of visual information, it would not affect the response latency.

Overall, the results highlight the efficiency advantages of Flash-VStream in terms of both response latency and GPU memory consumption, making it a competitive choice for real-time long video understanding tasks.

D. Ablation Study on Memory Structure

In Sec. 4.4 and Fig. 4, we initially explored the relationship between memory allocation strategy and pool ratio of CSM and DAM. Empirically, we found the best setting for these configurations under the fixed-budget constraint. In this section, we aim to answer the following questions:

Q1: How sensitive is the model performance to cluster numbers of CSM, i.e., N^{CSM}?

Q2: How sensitive is the model performance to key frame numbers of DAM, i.e., N^{DAM}?

As presented in Tab. 2, we conduct two groups of experi-

ID	Memory Component Settings					Evaluation Results			
	CSM	DAM	$\text{N}_{\text{Vtokens}}$	CSM Size	DAM Size	EgoSchema	MVBench	Video-MME(w/o)	Average
	✓	✓	19200	60 × 64	60 × 256	68.6	65.5	61.2	65.1
	✓	✓	15360	60 × 64	45 × 256	68.3	65.3	61.0	64.9
①	✓	✓	11520	60 × 64	30 × 256	68.2	65.4	61.2	64.9
	✓	✓	7680	60 × 64	15 × 256	67.5	64.9	60.8	64.4
③	✓	✗	3840	60 × 64	0	66.8	64.0	60.1	63.6
	✓	✗	5760	90 × 64	0	66.6	63.9	61.0	63.8
③	✓	✗	3840	60 × 64	0	66.8	64.0	60.1	63.6
	✓	✗	1920	30 × 64	0	65.7	63.6	58.8	62.7
	✓	✗	960	15 × 64	0	63.0	63.0	58.3	61.5

Table 2. **Ablation study of memory structure configurations.** We investigate the model's sensitivity to cluster numbers of CSM and key frame numbers of DAM.

Score	0	1	2	3	4	5	Total	Average Score
Right	8	0	26	111	1916	2732	4793	4.53
Wrong	355	290	1712	82	82	686	3207	2.41
Total	363	290	1738	193	1998	3418	8000	3.68

Table 3. **Score distribution of a GPT-3.5-based evaluation.** We tested Qwen2-VL-7b on ActivityNet-QA benchmark, using GPT-3.5-turbo-0125 for evaluation. It is observed that many wrong predictions are assigned with a high score "5", leading to a biased result.

ments to investigate the model's sensitivity to memory structure configurations, i.e., memory sizes N^{CSM} and N^{DAM}. In each group, we compare different memory size choices to the baseline row ① and row ③ in Table 4. The results show a scaling trend of accuracy with different memory sizes. Therefore, the results of grid search experiment illustrated in Fig. 4 are reasonable.

E. Case Study

In this section, we conduct a case study to provide a comprehensive understanding of the performance of models. This study presents a series of visual cases involving various types of videos, each accompanied by a specific question and multiple-choice options to evaluate the performance of three different models: Qwen2-VL [8], LLaVA-OV [4], and the proposed Flash-VStream.

Figs. 2 to 7 present different genres of videos, including documentaries, cartoons, commercials, sports programs and tutorial videos. As shown in these cases, Flash-VStream exhibits strong understanding capabilities in object recognition, action recognition, action reasoning, temporal reasoning, object counting and object reasoning.

F. Limitations

F.1. Fail Case Analysis

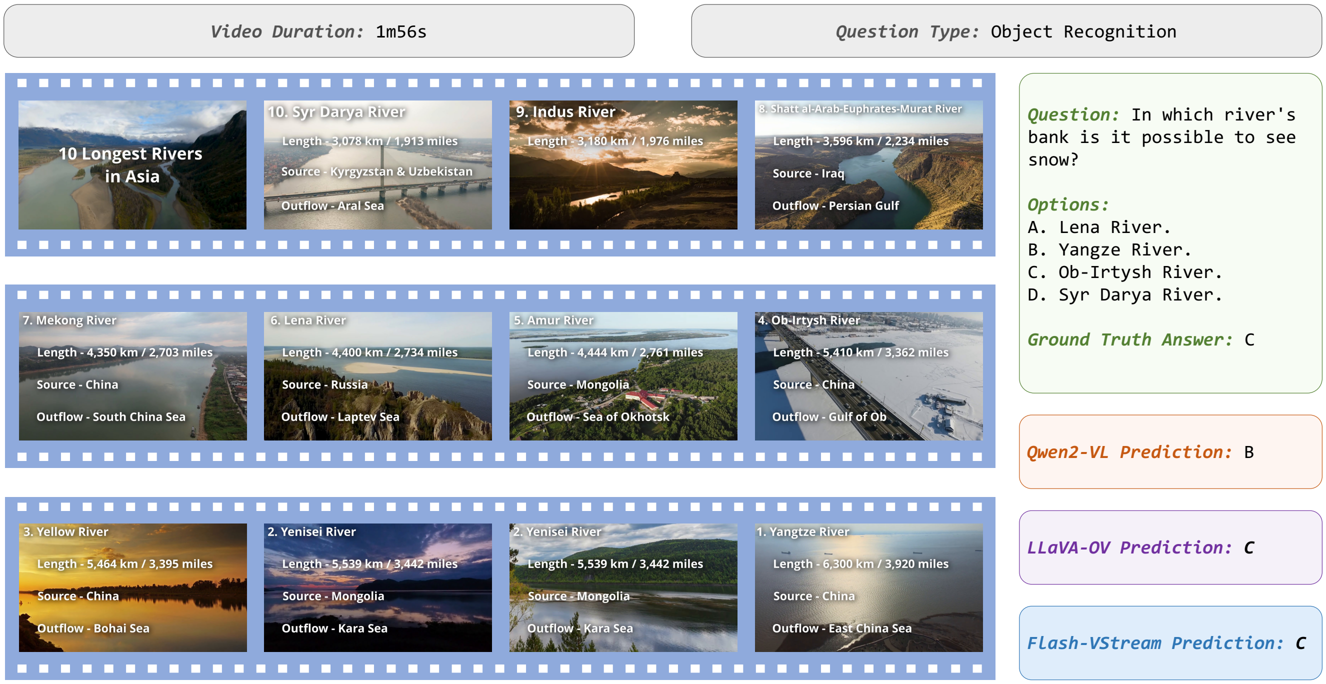
Flash-VStream may produce incorrect predictions in certain scenarios, such as text-intensive long videos (see Fig. 8) and long videos with rapid scene changes (see Fig. 9). We suggest that these heavily edited video have a different information density distribution compared to native videos, making efficient timing modeling more difficult.

F.2. GPT-3.5-based Metric for Open-ended VQA

It is worth noting that some previous works [2, 5, 7] follow Video-ChatGPT [6] to test models on open-ended VQA benchmarks [9, 10] based on GPT-3.5-based judgment (GPT accuracy and GPT score). However, we notice that these metrics are highly unstable and prone to bias, so we try to avoid evaluating models on these *LLM-as-a-judge* benchmarks. Since GPT APIs are proprietary and upgrade over time, this evaluation approach lacks reliability, stability and reproducibility [3]. Furthermore, the evaluation can be disturbed by the hallucination of GPT, leading to a biased evaluation result [7]. As presented in Tab. 3, there is always a discrepancy between the distribution of GPT accuracy and GPT score. Therefore, it is still challenging to benchmark the open-ended VQA ability of MLLMs.

G. Future Work

Future work could focus on enhancing the models' ability to understand edited videos with intensive text or rapid scene transitions, while maintaining the overall efficiency. Another interesting direction for future work would be to investigate reliable evaluation methods for open-ended VQA. Additionally, the techniques developed in this study could be adapted for use in other fields such as robotics and surveillance systems. We hope that our work will inspire further innovations and improvements in these fields, ultimately leading to more intelligent and versatile systems.



Video Duration: 1m56s

Question Type: Object Recognition

Question: In which river's bank is it possible to see snow?

Options:
A. Lena River.
B. Yangze River.
C. Ob-Irtysh River.
D. Syr Darya River.

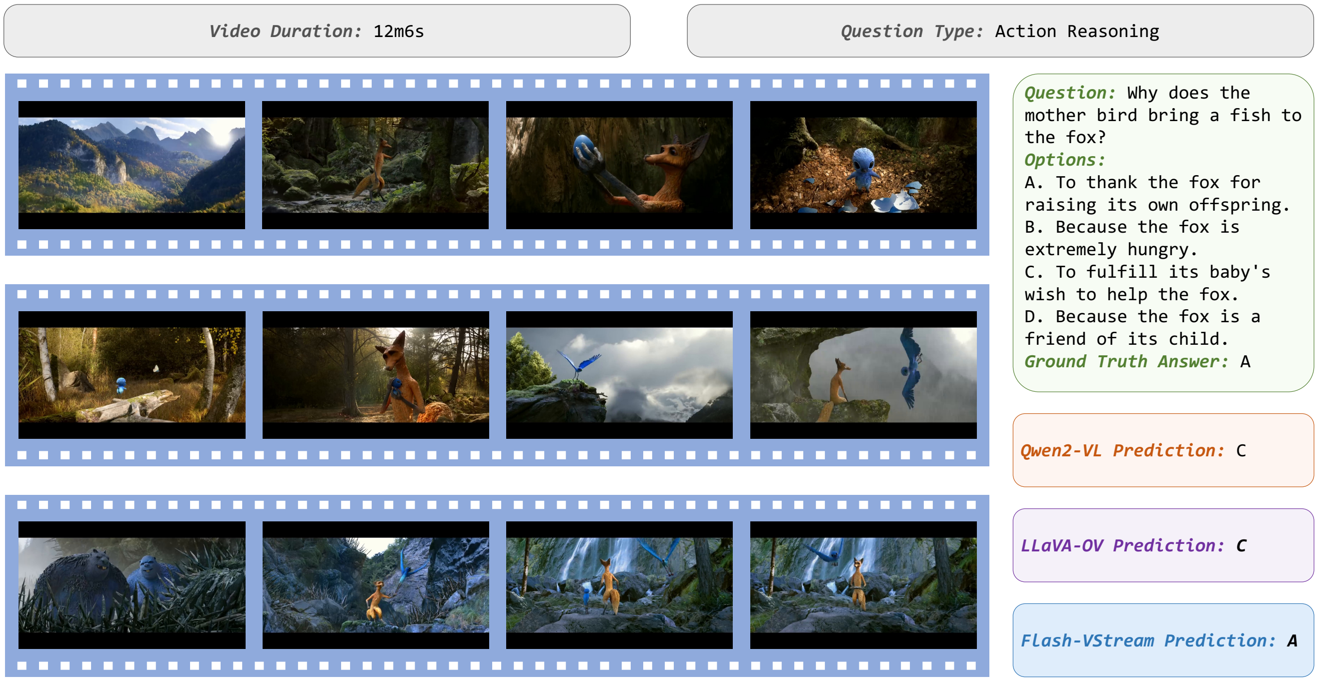
Ground Truth Answer: C

Qwen2-VL Prediction: B

LLaVA-OV Prediction: *C*

Flash-VStream Prediction: *C*

Figure 2. **Case Study.** This figure presents a case study on documentary video about the 10 longest rivers in Asia, highlighting their lengths, sources, and outflows. The study includes a question regarding the possibility of seeing snow on the banks of these rivers, with multiple-choice options provided. The ground truth answer is indicated, along with the predictions from three different models: Qwen2-VL, LLaVA-OV, and Flash-VStream.



Video Duration: 12m6s

Question Type: Action Reasoning

Question: Why does the mother bird bring a fish to the fox?
Options:
A. To thank the fox for raising its own offspring.
B. Because the fox is extremely hungry.
C. To fulfill its baby's wish to help the fox.
D. Because the fox is a friend of its child.
Ground Truth Answer: A

Qwen2-VL Prediction: C

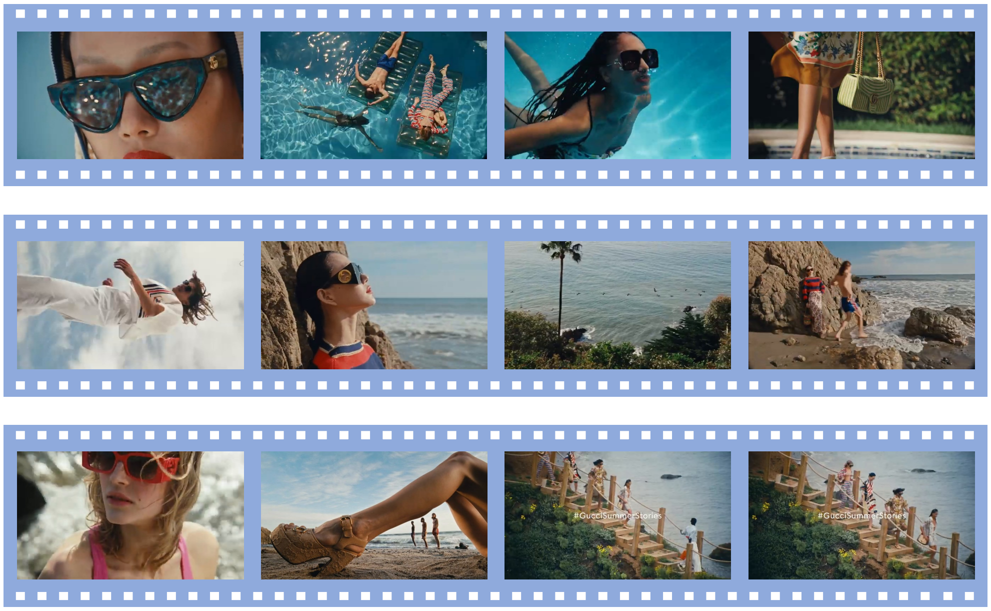
LLaVA-OV Prediction: *C*

Flash-VStream Prediction: *A*

Figure 3. **Case Study.** This figure presents a case study involving a cartoon video depicting a mother bird bringing a fish to a fox. The study includes a question about the reason behind this action, with multiple-choice options provided. The ground truth answer is indicated, along with the predictions from three different models: Qwen2-VL, LLaVA-OV, and Flash-VStream.

Video Duration: 1m0s

Question Type: Counting Problem



Question: At the end of the video, how many people are there on the staircase along a coastal hillside?

Options:
A. 3.
B. 2.
C. 4.
D. 5.

Ground Truth Answer: ***D***

Qwen2-VL Prediction: ***A***

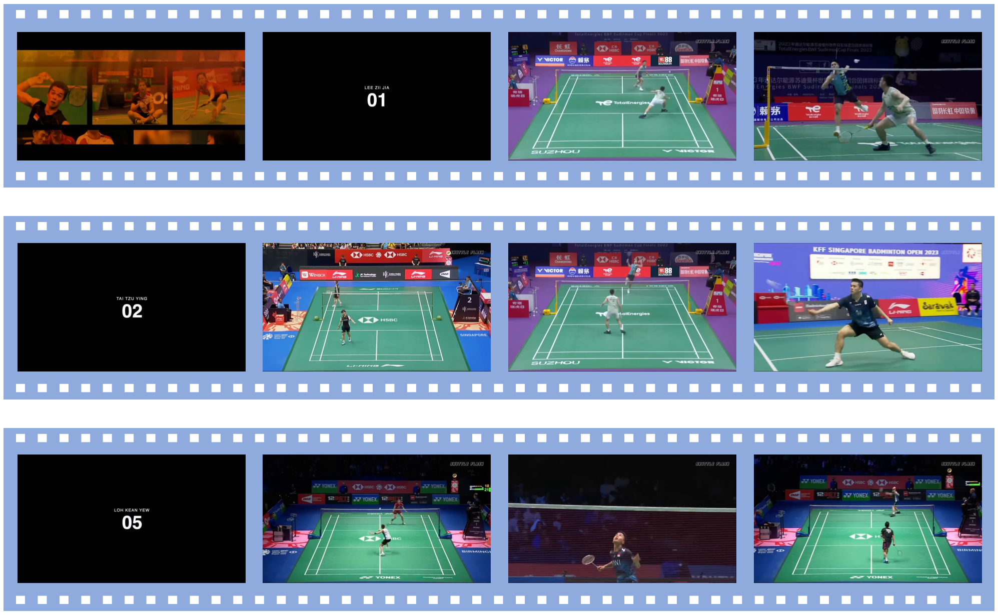
LLaVA-OV Prediction: ***C***

Flash-VStream Prediction: ***D***

Figure 4. **Case Study.** This figure presents a case study involving an advertising video, depicting various scenes including people by the pool, on the beach, and along a coastal hillside. The study includes a question about the number of people on the staircase at the end of the video, with multiple-choice options provided. The ground truth answer is indicated, along with the predictions from three different models: Qwen2-VL, LLaVA-OV, and Flash-VStream.

Video Duration: 6m21s

Question Type: Spatial Perception



Question: Where is the first match held?

Options:
A. SINGAPORE.
B. SUZHOU.
C. Sudirman.
D. HANGZHOU.

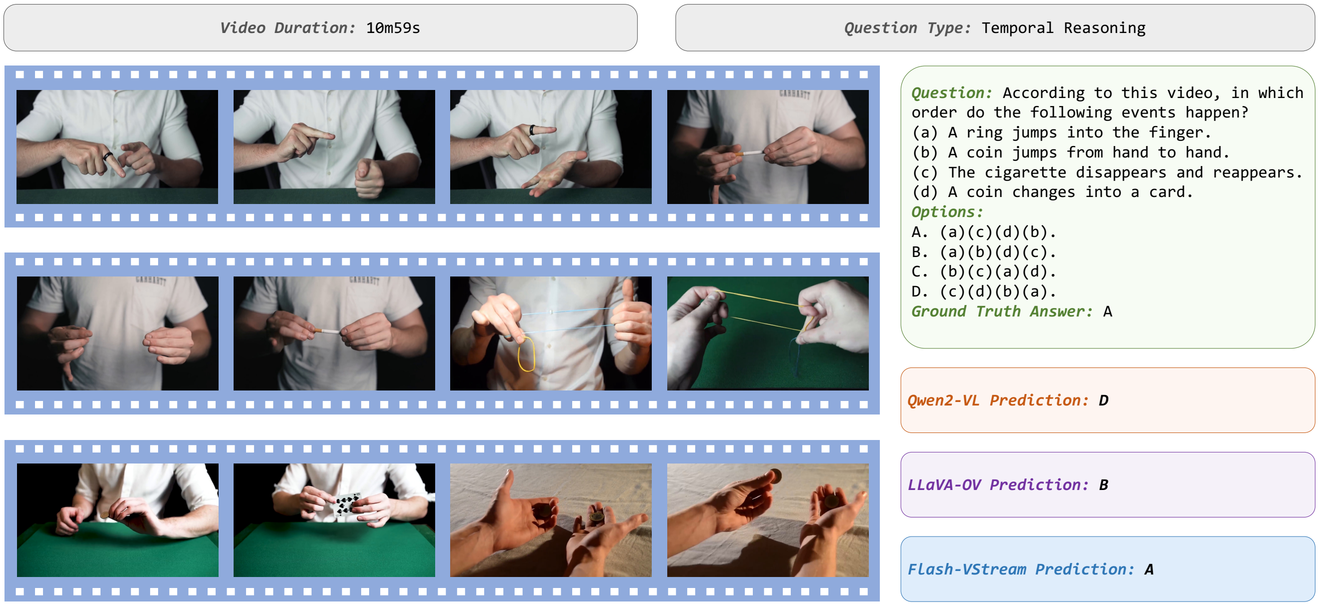
Ground Truth Answer: ***B***

Qwen2-VL Prediction: ***A***

LLaVA-OV Prediction: ***C***

Flash-VStream Prediction: ***B***

Figure 5. **Case Study.** This figure presents a case study involving a sports documentary video of badminton tournaments, depicting various matches and players. The study includes a question about the location of the first match, with multiple-choice options provided. The ground truth answer is indicated, along with the predictions from three different models: Qwen2-VL, LLaVA-OV, and Flash-VStream.



Video Duration: 10m59s

Question Type: Temporal Reasoning

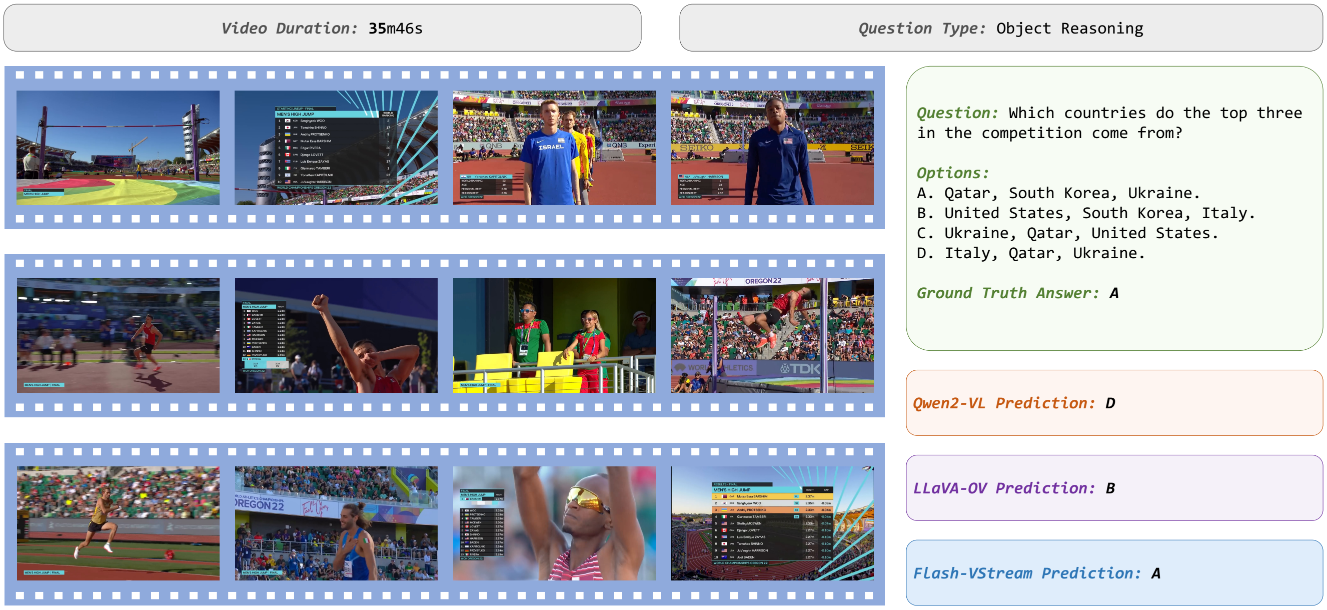
Question: According to this video, in which order do the following events happen?
(a) A ring jumps into the finger.
(b) A coin jumps from hand to hand.
(c) The cigarette disappears and reappears.
(d) A coin changes into a card.
Options:
A. (a)(c)(d)(b).
B. (a)(b)(d)(c).
C. (b)(c)(a)(d).
D. (c)(d)(b)(a).
Ground Truth Answer: A

Qwen2-VL Prediction: ***D***

LLaVA-OV Prediction: ***B***

Flash-VStream Prediction: ***A***

Figure 6. **Case Study.** This figure presents a case study involving a tutorial video depicting various magic tricks. The study includes a question about the order of events in the video, with multiple-choice options provided. The ground truth answer is indicated, along with the predictions from three different models: Qwen2-VL, LLaVA-OV, and Flash-VStream.



Video Duration: 35m46s

Question Type: Object Reasoning

Question: Which countries do the top three in the competition come from?

Options:
A. Qatar, South Korea, Ukraine.
B. United States, South Korea, Italy.
C. Ukraine, Qatar, United States.
D. Italy, Qatar, Ukraine.

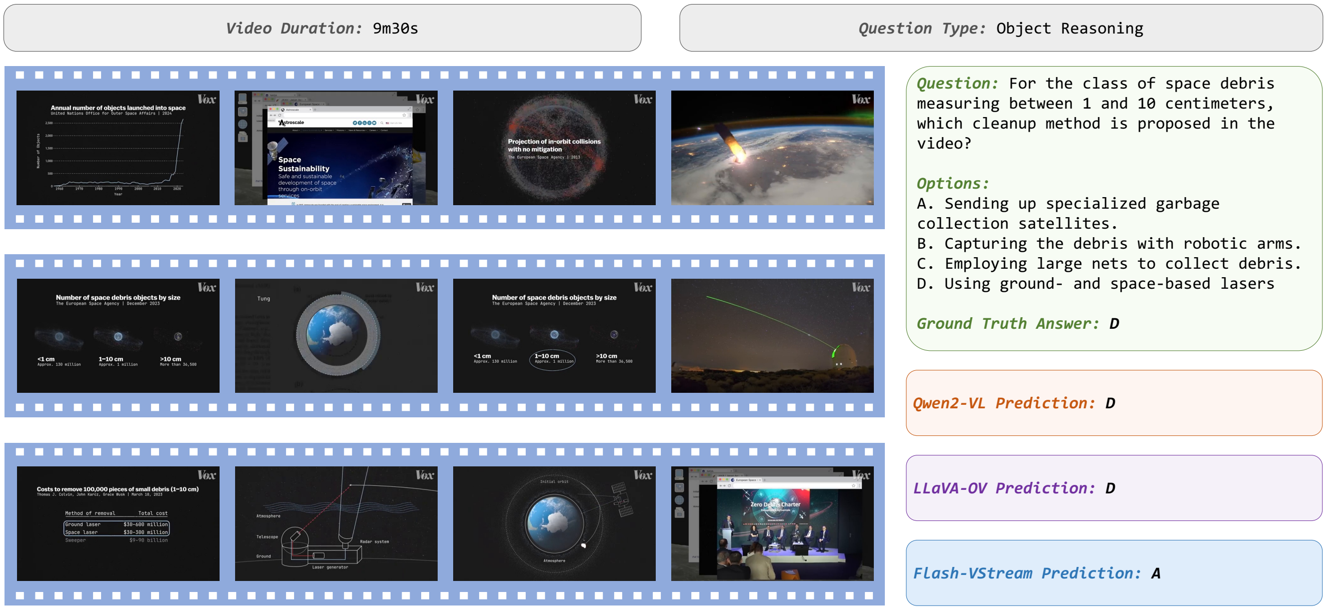
Ground Truth Answer: ***A***

Qwen2-VL Prediction: ***D***

LLaVA-OV Prediction: ***B***

Flash-VStream Prediction: ***A***

Figure 7. **Case Study.** This figure presents a case study involving a sports video from a high jump competition, depicting various athletes and their performances. The video frames capture moments of intense competition, showcasing the athletes' skills and determination as they strive to achieve their best performances. The analysis aims to evaluate the models' ability to accurately interpret and predict the outcomes based on visual and contextual cues from the video. The study includes a question about the countries of the top three athletes in the competition, with multiple-choice options provided. The ground truth answer is indicated, along with the predictions from three different models: Qwen2-VL, LLaVA-OV, and Flash-VStream.



Video Duration: 9m30s

Question Type: Object Reasoning

Question: For the class of space debris measuring between 1 and 10 centimeters, which cleanup method is proposed in the video?

Options:
A. Sending up specialized garbage collection satellites.
B. Capturing the debris with robotic arms.
C. Employing large nets to collect debris.
D. Using ground- and space-based lasers

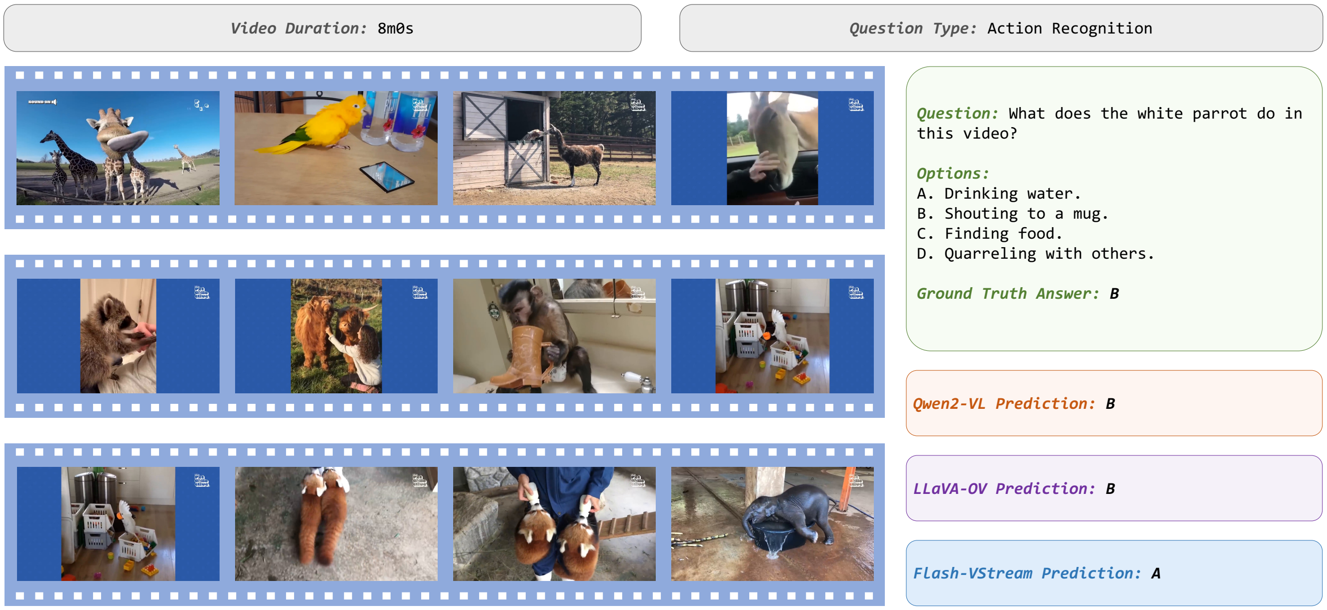
Ground Truth Answer: D

Qwen2-VL Prediction: D

LLaVA-OV Prediction: D

Flash-VStream Prediction: A

Figure 8. **Fail Case Analysis.** This figure presents a case study involving a video on space debris and proposed cleanup methods. The video frames illustrate various statistics and methods related to space debris, highlighting the challenges and potential solutions for mitigating the growing problem of space junk. The study includes a question about the recommended method for cleaning up space debris measuring between 1 and 10 centimeters, with multiple-choice options provided. The ground truth answer is indicated, along with the predictions from three different models: Qwen2-VL, LLaVA-OV, and Flash-VStream.



Video Duration: 8m0s

Question Type: Action Recognition

Question: What does the white parrot do in this video?

Options:
A. Drinking water.
B. Shouting to a mug.
C. Finding food.
D. Quarreling with others.

Ground Truth Answer: B

Qwen2-VL Prediction: B

LLaVA-OV Prediction: B

Flash-VStream Prediction: A

Figure 9. **Fail Case Analysis.** This figure presents a case study involving a video showing various animals and their behaviors. The video frames capture different moments of animal interactions and activities, highlighting the diverse behaviors exhibited by the animals. The study includes a question about the specific action of a white parrot in the video, with multiple-choice options provided. The ground truth answer is indicated, along with the predictions from three different models: Qwen2-VL, LLaVA-OV, and Flash-VStream.

References

[1] Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. In *ICLR*, 2024. 2

[2] Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *CVPR*, pages 13700–13710, 2024. 2, 3

[3] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *CVPR*, pages 13299–13308, 2024. 3

[4] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 2, 3

[5] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *ECCV*, pages 323–340. Springer, 2025. 2, 3

[6] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *ACL*, pages 12585–12602, 2024. 3

[7] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. Moviechat: From dense token to sparse memory for long video understanding. In *CVPR*, pages 18221–18232, 2024. 2, 3

[8] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 2, 3

[9] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *CVPR*, pages 9777–9786, 2021. 3

[10] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *AAAI*, pages 9127–9134, 2019. 3

[11] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024. 2