# SAGE: Training Smart Any-Horizon Agents for Long Video Reasoning with Reinforcement Learning
# Supplementary Material

Jitesh Jain[1,2*] Jialuo Li[1] Zixian Ma[2,3] Jieyu Zhang[2,3] Chris Dongjoo Kim[2] Sangho Lee[2]
Rohun Tripathi[2] Tanmay Gupta[2] Christopher Clark[2†] Humphrey Shi[1†]
[1]SHI Labs @ Georgia Tech [2]Allen AI [3]University of Washington

https://github.com/allenai/SAGE

| train strategy | train mode | eval mode | mcq | open-ended | overall |
|---|---|---|---|---|---|
| Qwen3-VL-4B-Instruct | | DIRECT | 75.8 | 51.5 | 62.7 |
| Qwen3-VL-4B-Thinking | | DIRECT | 75.3 | 48.6 | 60.1 |
| SFT | DIRECT | DIRECT | **83.2** | 51.1 | 65.8 |
| SFT + RL | DIRECT | DIRECT | 83.0 | 52.0 | 66.3 |
| SFT (ours) | AGENT | AGENT | 77.3 | 53.7 | 64.6 |
| SFT + RL (ours) | AGENT | AGENT | 81.3 | **57.4** | **68.4** |

Table 1. **Training Mode.** Our AGENT system performs better than the DIRECT baseline, with RL playing a critical role in the former's success, specifically on open-ended problems.

In this supplementary material, we first present additional ablations in Sec. 1, including the effect of video input on SAGE-Bench, the importance of the cold-start SFT stage, and the impact of varying $N_{max}$ during evaluation. Secondly, we provide qualitative examples from our SAGE-Bench in Sec. 3 along with a comparison to existing benchmarks. Next, we list all the system prompts used in our work in Sec. 4. Lastly, we list qualitative examples from the QnA pair generation pipeline in Sec. 5. Unless mentioned otherwise, we use the Qwen3-VL-8B-Instruct-based SAGE-MM for all experiments in this supplementary material and report results after RL on SAGE-Bench.

## 1. Additional Ablations

**Training Mode.** In Tab. 1, we finetune a Qwen3-VL-4B-Instruct model on the synthetic QnA pairs with DIRECT answering mode under the same data setting. We observe that our AGENT training recipe outperforms the direct baseline, underscoring the effectiveness of our approach. Specifically, while training the DIRECT baseline with SFT, we supervise the model with only the correct final answer and not the tool call actions. During RL, we use only the accuracy reward to train the DIRECT baseline.

*Work done during JJ's internship at Allen AI. †Equal advising.

| system | SAGE-MM | single-turn | | multi-turn | | overall |
|---|---|---|---|---|---|---|
| | Qwen3-VL-8B-Instruct *(base)* | count | acc. | count | acc. | acc. |
| SAGE-Flash | Gemini-2.5-Flash (expert) | 859 | 76.9 | 885 | 66.0 | 71.3 |
| **SAGE** | [+SFT] (ours) | 706 | 79.0 | 1038 | 53.7 | 64.6 |
| **SAGE** | [+SFT] [+RL] (ours) | 948 | 79.6 | 796 | 54.3 | 68.0 |
| **SAGE-Flash** | [+SFT] [+RL] (ours) | 940 | 78.8 | 804 | 63.4 | 71.8 |

Table 2. **Any-Horizon Reasoning.** RL refines the tool's overcalling behavior of the SFT model, resulting in a distribution closer to the expert Gemini-2.5-Flash and thus, improved performance.

| method | video | overall | both | verbal | visual |
|---|---|---|---|---|---|
| Qwen3-VL-8B-Instruct | ✓ | **64.9** | **72.8** | 68.7 | **61.9** |
| Qwen3-VL-8B-Instruct | ✗ | 42.1 | 58.6 | 70.9 | 33.6 |
| **SAGE** (ours) | ✓ | **68.0** | **75.4** | **82.8** | **64.0** |
| **SAGE** (ours) | ✗ | 41.0 | 48.0 | 35.8 | 39.3 |

Table 3. **Importance of the Video Input.** Access to the video is critical for good performance on SAGE-Bench.

**Any-Horizon Reasoning.** A core aspect of SAGE's design is to enable any-horizon reasoning, *i.e.*, it is adept at multi-turn reasoning and also directly outputting an answer in a single step. As shown in Tab. 2, our SFT model, distilled from the expert Gemini-2.5-Flash, inherits strong single-turn ability but tends to show signs of overcalling tools. Incorporating RL further refines this behavior while improving single-turn and multi-turn accuracies.

**Importance of the Video Input.** Although we build SAGE-Bench using questions strictly disjoint from the training set, some videos in SAGE-Bench overlap with those seen during training. Therefore, it is essential to assess potential memorization. A natural test is to evaluate whether the model produces correct answers without access

| SAGE-MM | system | eval mode | overall | mcq | open-ended |
|---|---|---|---|---|---|
| *Qwen3-VL-8B-Instruct* | | | | | |
| [+SFT] | Qwen3-VL | DIRECT | 63.6 | 78.4 | 50.9 |
| | SAGE | AGENT | 63.9 | 77.4 | 52.4 |
| | SAGE-Flash | AGENT | **70.5** | 81.0 | 61.5 |
| [+SFT] [+RL] | Qwen3-VL | DIRECT | 69.8 | 84.0 | 57.6 |
| | SAGE | AGENT | 68.0 | 82.6 | 55.6 |
| | SAGE-Flash | AGENT | **71.8** | 82.8 | 62.4 |
| *Molmo2-8B* | | | | | |
| [+SFT] | Molmo2 | DIRECT | 55.7 | 68.5 | 44.9 |
| | SAGE | AGENT | 63.3 | 75.6 | 52.8 |
| | SAGE-Flash | AGENT | **69.8** | 79.9 | 61.3 |
| [+SFT] [+RL] | Molmo2 | DIRECT | 61.0 | 71.7 | 51.9 |
| | SAGE | AGENT | 66.1 | 78.8 | 55.2 |
| | SAGE-Flash | AGENT | **67.8** | 79.3 | 58.1 |

Table 4. **Eval Mode.** We find that for the trained SAGE-MM, AGENT mode during inference outperforms the DIRECT mode. All the models are trained under the AGENT paradigm

to the video. We find no evidence of memorization: performance drops by 27%, similar to the drop observed for the base Qwen3-VL-8B-Instruct model, as shown in Tab. 3, underscoring the validity of our approach and findings.

**Eval Mode.** In Tab. 4, we analyze the effect of eval mode with the trained Qwen3-VL-8B-Instruct SAGE-MM during inference. We find that the AGENT mode performs better than DIRECT mode. Surprisingly, the Qwen3-VL [10] based model shows much better performance than the Molmo2 [4] one with DIRECT eval mode which could be attributed to the two model families' different abilities to learn information directly since all the models are trained under the AGENT paradigm.

**Importance of SFT.** Our SAGE is designed so that any MLLM with function-calling capabilities can be used as the SAGE-MM. In Tab. 5, we evaluate the base Qwen3-VL [10] models as SAGE-MM, without any finetuning. We observe that Qwen3-VL-4B-Instruct is not an effective orchestrator: it rarely engages in multi-turn reasoning and attains low accuracy, indicating that SFT is essential before applying RL.

Interestingly, the base Qwen3-VL-8B-Instruct model behaves differently. It is a noticeably stronger function caller, demonstrating a more reasonable balance between single-turn and multi-turn reasoning. This motivates us to apply RL directly on top of the base model to assess the importance of SFT for the 8B variant. Surprisingly, RL without SFT fails, *i.e.*, the model collapses to single-turn reasoning. We hypothesize that this is due to the base model's training objective, which strongly biases it toward directly producing final answers, making SFT necessary to incentivize [12] any-horizon reasoning during RL. While it is possible that a heavily engineered RL recipe could overcome this, we do not pursue this direction, as SFT is far simpler and cheaper than extensive hyperparameter tuning during RL.

**#Turns v/s Video Duration.** In Tab. 6, we report the average number of reasoning turns across all samples grouped by video duration buckets. We observe a gradual increase in the number of turns as video length increases, indicating that SAGE naturally adapts its trajectory length to the temporal horizon of the input. Shorter videos lead to shorter reasoning trajectories, whereas longer videos elicit more extended ones, aligned with our design objective of instilling any-horizon reasoning into the system.

**Effect of $N_{max}$.** We study the effect of varying $N_{max}$ during evaluation in Tab. 7. We find that setting $N_{max} = 11$ achieves high accuracy while keeping the number of unanswered samples low, with only minimal gains from further increases in $N_{max}$. This demonstrates the effectiveness of our RL recipe in enforcing answer prediction within an 11-step reasoning horizon.

**Variance on SAGE-Bench.** We analyze the variance in performance on SAGE-Bench across five different runs (with temperature of 1.0) in Tab. 8. We find a low standard deviation of 0.22 for the base Qwen3-VL-8B-Instruct model. The low variance indicates that the performance improvements from our SAGE are statistically significant.

**Importance of Supported Tools.** We ablate the contribution of each tool in Tab. 9. Dropping the *transcribe-speech*, *extract-video-parts*, and *analyze* tools leads to the most significant performance decline, highlighting their fundamental role in long-video reasoning. In contrast, removing the *ground-event* tool results in only a minor drop, likely due to the tool's inherent inaccuracy. This observation underscores the need for developing better temporal grounding modules.

**Evaluation on Video-MME and Video-MMMU.** We evaluate SAGE on two widely-used benchmarks: Video-MME [7] and Video-MMMU [8] in Tab. 10. We find that our SAGE-Flash outperforms the baselines on Video-MMMU, demonstrating generalization to knowledge acquisition from videos. Interestingly, we see a drop in performance on Video-MME attributed to its general perception-centric nature where `ground-event/extract-video-parts` tools harm performance, unlike on Video-MMMU and SAGE-Bench (Tab. 9 in the main paper). Notably, SAGE-Flash convincingly outperforms baselines including Video-R1 [6] on Video-MMMU.

**Per-Tool Accuracy.** We report per-tool accuracy in Tab. 11, evaluating performance when only a single tool is available. We observe that `extract-video-parts/ground-event` perform the worst, attributed to their dependence on tools like `analyze/extract-video-parts` for local segment processing. The best individual tool performance comes

| SAGE-MM | single-turn | | multi-turn | | overall |
|---|---|---|---|---|---|
| | count | acc. | count | acc. | acc. |
| Qwen3-VL-4B-Instruct | 1345 | 54.6 | 399 | 52.8 | 54.5 |
| Qwen3-VL-4B-Instruct [+SFT] **(ours)** | 691 | 79.5 | 1045 | 54.7 | 64.6 |
| Qwen3-VL-4B-Instruct [+SFT] [+RL] **(ours)** | 832 | 80.5 | 912 | 57.3 | **68.4** |
| Qwen3-VL-8B-Instruct | 802 | 79.5 | 942 | 53.7 | 63.2 |
| Qwen3-VL-8B-Instruct [+RL] | 1727 | 56.9 | 17 | 23.6 | 56.6 |
| Qwen3-VL-8B-Instruct [+SFT] **(ours)** | 706 | 79.0 | 1038 | 53.7 | 63.9 |
| Qwen3-VL-8B-Instruct [+SFT] [+RL] **(ours)** | 948 | 79.6 | 796 | 54.3 | **68.0** |

Table 5. **Importance of SFT.** The cold-start SFT stage is necessary to incentivize multi-turn reasoning during RL.

| SAGE-MM | 0-60 | 60-180 | 180-300 | 300-600 | 600-1200 | 1200-2400 | 2400+ |
|---|---|---|---|---|---|---|---|
| Qwen3-VL-8B-Instruct [+SFT] | 2.00 | 2.23 | 2.05 | 2.63 | 3.02 | 3.50 | 3.54 |
| Qwen3-VL-8B-Instruct [+SFT] [+RL] | 1.74 | 1.81 | 1.83 | 2.18 | 2.49 | 2.89 | 2.77 |

Table 6. **#Turns v/s Video Duration.** The average number of reasoning turns grows gradually with an increase in video duration, demonstrating our SAGE's any-horizon nature.

| $N_{\max} \rightarrow$ | | 1 | | 2 | | 3 | | 6 | | 11 (default) | | 13 | | 16 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| system | SAGE-MM | acc. | no ans. | acc. | no ans. | acc. | no ans. | acc. | no ans. | acc. | no ans. | acc. | no ans. | acc. | no ans. |
| SAGE | Qwen3-VL-8B-Instruct [+SFT] | 33.8 | 60.8 | 46.0 | 45.0 | 57.0 | 23.4 | 62.1 | 5.5 | 63.9 | 3.3 | 63.8 | 3.5 | 64.6 | 3.0 |
| SAGE | Qwen3-VL-8B-Instruct [+SFT] [+RL] | 43.4 | 46.8 | 56.7 | 30.4 | 62.8 | 20.2 | 66.6 | 3.9 | 68.0 | 1.3 | 67.8 | 1.1 | 67.9 | 1.1 |
| SAGE-Flash | Qwen3-VL-8B-Instruct [+SFT] [+RL] | 43.5 | 47.1 | 57.1 | 30.1 | 63.2 | 21.2 | 70.7 | 6.3 | 71.8 | 3.3 | 72.2 | 2.9 | 71.9 | 2.9 |

Table 7. **Effect of $N_{max}$.** Limiting the total number of turns to 11 is optimal as our RL recipe enforces the ability to produce an answer in as many turns. *no ans.* denotes the percentage of samples where an answer could not be produced. *acc.* denotes the accuracy score.

| run-1 | run-2 | run-3 | run-4 | run-5 | mean | std |
|---|---|---|---|---|---|---|
| 64.9 | 64.6 | 64.9 | 65.2 | 65.1 | **64.9** | **0.22** |

Table 8. **Variance on SAGE-Bench.** We find a low standard deviation of 0.22 across five different runs with Qwen3-VL-8B-Instruct with temperature set to 1.0.

| | overall | both | verbal | visual |
|---|---|---|---|---|
| **SAGE** (ours) | **68.0** | **75.4** | **82.8** | 64.0 |
| w/o ground-event | 67.3 | 72.3 | 79.9 | **64.3** |
| w/o web-search/parse-website | 65.5 | 70.1 | 80.6 | 62.4 |
| w/o analyze | 63.4 | 70.6 | 80.6 | 59.1 |
| w/o extract-video-parts | 63.0 | 70.8 | 79.9 | 58.6 |
| w/o transcribe-speech | 62.5 | 66.8 | 46.3 | 62.9 |

Table 9. **Dropping Tools during inference.** All tools are critical to the success of SAGE as a system, with the extract-video-parts and transcribe-speech being the most important ones for answering the visual and verbal/both questions, respectively, as expected.

from `transcribe-speech`, highlighting the importance of speech information for long video understanding.

**Qualitative Failure Analysis.** We observe a few failure patterns of SAGE. First, SAGE may overcall tools when a tool invocation fails, resulting in retries that consume reasoning steps. Second, inaccurate temporal grounding can result in missing video segments, leading to incomplete context. Lastly, misleading transcript or web information can result in incorrect answers.

| SAGE-MM | | Video-MMMU | Video-MME |
|---|---|---|---|
| Qwen2.5-VL-7B-Instruct [1] | N/A | 57.5 | 63.6 |
| Qwen3-VL-8B-Instruct [10] | N/A | 65.3 | **66.8** |
| Video-R1-7B [6] | N/A | 61.5 | 61.2 |
| **SAGE-Flash** (ours) | Qwen3-VL-8B-Instruct [+SFT] | 66.9 | 59.4 |
| **SAGE-Flash** (ours) | Qwen3-VL-8B-Instruct [+SFT] [+RL] | **68.1** | 63.5 |
| - w/o ground-event | | 65.8 | 65.6 |
| - w/o ground-event & extract-video-parts | | 61.8 | 66.2 |

Table 10. **Video-MMMU & Video-MME (w/o subs) evaluation.**

| | | | | | | |
|---|---|---|---|---|---|---|
| **all-tools** | **71.8** | transcribe-speech | 61.1 | web-search/parse-website | | 58.6 |
| analyze | 58.6 | extract-video-parts | 50.2 | ground-event | 50.3 | no-tools | 53.0 |

Table 11. **Per-tool performance comparison on SAGE-Bench.**

## 2. Eval Runtime

In Tab. 12, we compare the accuracy score and inference runtime per sample for our SAGE to other existing DIRECT [6, 11] and AGENT [2, 5, 15] baselines and various frame-input setups of the baseline Qwen3-VL-8B-Instruct [10]. We observe that although the runtime of our SAGE is comparable to using 512 frames as inputs to

| method | mode | #frames | acc. | runtime (*sec/sample*) |
|---|---|---|---|---|
| Qwen3-VL-8B-Instruct | DIRECT | 16 | 55.7 | 0.8 |
| | | 32 | 59.3 | 1.1 |
| | | 64 | 62.3 | 2.3 |
| | | 128 | 64.9 | 3.6 |
| | | 256 | 66.1 | 5.7 |
| | | 512 | 65.9 | 7.8 |
| | | 1024 | 62.5 | 18.3 |
| | | 1536 | 60.8 | 27.5 |
| VideoRFT-7B [11] | DIRECT | 128 | 55.3 | 7.2 |
| Video-R1-7B [6] | DIRECT | 128 | 57.6 | 7.3 |
| VideoMind-7B [9] | AGENT | — | 50.0 | 24.7 |
| LVAagent [2] | AGENT | — | 49.7 | 92.9 |
| VideoChat-R1.5-7B [14] | AGENT | — | 54.8 | 132.1 |
| VideoExplorer-7B [15] | AGENT | — | 50.1 | 137.7 |
| VideoAgent [5] | AGENT | — | 42.0 | 1445.0 |
| **SAGE** | AGENT | — | **68.0** | **8.6** |

Table 12. **Eval Runtime.** Our SAGE shows a good performance-efficiency tradeoff owing to its any-horizon reasoning nature.

Qwen3-VL-8B-Instruct, it shows far superior performance, while only being slower by about 1 second compared to other thinking DIRECT baselines. Moreover, our system is almost 3 times quicker than VideoMind [9], the quickest AGENT baseline, demonstrating the superiority of our system design and training recipe for practical applications over existing systems.

The lower runtime of our framework compared to the AGENT baselines is primarily due to the baselines' system design, which involves heavy video preprocessing and excessive recurrent model calls. Notably, VideoAgent [5] is slowed by a mandatory preprocessing phase in which every 2-second subclip undergoes multi-model analysis for metadata extraction, making it super slow for long videos. Similarly, VideoExplorer [15] suffers from both an initial 30-second preprocessing delay, arising from dividing the video into multiple subclips for embedding computation, and an inference process involving multiple retrieval steps. Finally, VideoMind [9] inherently requires more model invocations. This increase can be traced to the system design, which requires repetitive invocations of the verifier module. The verifier is executed multiple times, once for each of the top five potential segments generated by the preceding grounder module, which slows the system.

## 3. SAGE-Bench

Compared to existing video understanding benchmarks, SAGE-Bench demonstrates two distinct advantages:

**High-Quality Open-Ended Questions.** As illustrated in Fig. 1, existing popular benchmarks [3, 7, 8, 13, 16] rely purely on multiple-choice questions (MCQs). In contrast, SAGE-Bench utilizes open-ended questions with an unbounded answer space, aligning more closely with practical user situations.

**Dual Focus on Diagnostic and Practical Evaluation.** While AI systems are ultimately intended for real-world deployment, existing benchmarks often include diagnostic questions to gauge models' visual understanding, such as temporal ordering tasks in MLVU [16]. SAGE-Bench incorporates both diagnostic questions to test fundamental model capabilities and practical questions that users may have while watching entertainment videos, ensuring that our benchmark evaluates not only technical proficiency but also the model's utility in practical scenarios.

## 4. System Prompts

We provide information about the system prompts used for different purposes in this work below:

- QnA Pair Generation: Fig. 2.
- LLM-Judge Evaluation: Fig. 3.
- SAGE Stage-1 (Context VLM): Fig. 4.
- SAGE Stage-2 (Iterative Reasoner): Fig. 5.
- `ground-event` tool: Fig. 6.
- Reasonable Tool Step Reward Computation: Fig. 7.
- DIRECT baselines Evaluation: Fig. 8.

## 5. Qualitative Examples

**QnA Pairs.** We display some samples of the generated QnA pairs in Fig. 10 and Fig. 9.

**SAGE Any-Horizon Reasoning Trajectories.** We display qualitative examples to demonstrate the any-horizon reasoning abilities of SAGE in Fig. 11 (5 turns), Fig. 12 (2 turns), and Fig. 13 (single-turn).

## References


[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv*, 2025. 3

[2] Boyu Chen, Zhengrong Yue, Siran Chen, Zikang Wang, Yang Liu, Peng Li, and Yali Wang. Lvagent: Long video understanding by multi-round dynamical collaboration of mllm agents. *arXiv*, 2025. 3, 4

[3] Junhao Cheng, Yuying Ge, Teng Wang, Yixiao Ge, Jing Liao, and Ying Shan. Video-holmes: Can mllm think like holmes for complex video reasoning? *arXiv*, 2025. 4

[4] Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Rohun Tripathi, Sangho Lee, Mohammadreza Salehi, Jason Ren, Chris Dongjoo Kim, Yinuo Yang, Vincent Shao, Yue Yang, Weikai Huang, Ziqi Gao, Taira Anderson, Jianrui Zhang, Jitesh Jain, George Stoica, Winston Han, Ali Farhadi, and Ranjay Krishna. Molmo2: Open Weights and Data for Vision-Language Models with Video Understanding and


Figure 1. **Comparing SAGE-Bench to existing benchmarks.** SAGE-Bench contains samples covering both practical scenarios (IDs: *witFwlBjfLo, jjg4hWDFbmY, hh4prBn66Dc, 1SjmrYNHqiA*) and diagnostic cases. Representative examples for other benchmarks are sourced from their respective websites or papers.

Grounding. https://allenai.org/papers/molmo2, 2025. 2

[5] Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. Videoagent: A memory-augmented multimodal agent for video understanding. In *ECCV*, 2024. 3, 4

[6] Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. In *NeurIPS*, 2025. 2, 3, 4

[7] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis, 2025. 2, 4

[8] Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos, 2025. 2, 4

[9] Ye Liu, Kevin Qinghong Lin, Chang Wen Chen, and Mike Zheng Shou. Videomind: A chain-of-lora agent for long video reasoning. *arXiv*, 2025. 4

[10] Qwen Team. Qwen3-vl technical report. *arXiv*, 2025. 2, 3

[11] Qi Wang, Yanrui Yu, Ye Yuan, Rui Mao, and Tianfei Zhou. Videorft: Incentivizing video reasoning capability in mllms via reinforced fine-tuning. *arXiv*, 2025. 3, 4

[12] Zengzhi Wang, Fan Zhou, Xuefeng Li, and Pengfei Liu. Octothinker: Mid-training incentivizes reinforcement learning scaling. *arXiv*, 2025. 2

[13] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding, 2024. 4

[14] Ziang Yan, Xinhao Li, Yinan He, Zhengrong Yue, Xiangyu Zeng, Yali Wang, Yu Qiao, Limin Wang, and Yi Wang.

Videochat-r1.5: Visual test-time scaling to reinforce multi-modal reasoning by iterative perception. *arXiv*, 2025. 4

[15] Huaying Yuan, Zheng Liu, Junjie Zhou, Hongjin Qian, Yan Shu, Nicu Sebe, Ji-Rong Wen, and Zhicheng Dou. Think with videos for agentic long-video understanding. In *ICLR*, 2025. 3, 4

[16] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: Benchmarking multi-task long video understanding, 2025. 4

System Prompt to generate QnA pairs using Gemini-2.5-Flash

You are a specialized question generator. Your primary function is to generate 10–20 questions based on the provided video which can be upto 2 hours (7200 seconds) long.

- Pay attention to what modality information is needed to answer the question. You should generate questions that a viewer may be interested in and require visual, verbal, and or both in a balanced manner.
- You MUST give atleast four questions that cannot be answered with verbal information and require visual information.
- Also, it's okay to give questions that are not answerable from the video but can be answered with a web search.
- Generate a mix of open ended and multiple choice questions which are both hard and easy to answer. Err on the side of hard if you are unsure.

The duration of the video is <<<video_duration>>> seconds ( <<<timestamp_format>>> in HH:MM:SS format).
First think about the facts from the video and then generate questions about those. The questions could refer to the part of the video that spans across 10 seconds long but most MUST refer to the timeframes atleast a few minutes long.
Your timestamps MUST be in HH:MM:SS format.

**Output Format.** You MUST follow this format and MUST be between the <json> and </json> tags:

```
<json>
{
  "timestamp_format":"HH:MM:SS",
  "num_questions": <number of questions generated>,
  "questions": [
    {
      "index": <index_of_question_out_of_total_question>,
      "type": "type_of_question", // can be mcq or open_ended
      "difficulty": <difficulty_of_question>, // can be easy, medium, hard
      "difficulty_rationale": <why-this-difficulty>,
      "modality": <modality_of_question>, // can be visual, verbal, or both
      "modality_rationale": <why-this-modality>,
      "answer": <answer_text>, // answer for the question, if the type of question is
          mcq, then this is the text for the correct option, otherwise this is the
          answer text for the open ended question
      "question": <question_text>,
      "options": [ // if the type of question is mcq, then this is a list of options,
          otherwise this is null
        <option_1>, <option_2>, <option_3>, <option_4>, <option_5>, <option_6>
      ]
      "requires_web_search": <true | false>, // if the question requires a web search
          to be answered, then this is true, otherwise this is false
      "why_web_search": <reasoning for why web search is needed to answer the
          question>, // if the question requires a web search to be answered, then this
          is the reasoning for why web search is needed to answer the question,
          otherwise this is null
      "final_timestamp": <duration_of_the_video>, # HH:MM:SS
      "start_timestamp": <start_timestamp_of_question>, # HH:MM:SS
      "end_timestamp": <end_timestamp_of_question>, # HH:MM:SS
      "compute_percent_video_parsed": <think carefully and predict accurate
          percent_video_parsed, show calculation here>,
      "percent_video_parsed": <percentage_of_the_video_parsed_upto_this_question> #
          [(end_timestamp(seconds)/final_timestamp(seconds)) * 100] MUST go upto
          atleast 90 if not 100 for atleast one question
    },
    ...
  ]
}
</json>
```

This output will be converted to a JSON dict later on, you MUST use the correct syntax.

Figure 2. System Prompt to generate QnA pairs using Gemini-2.5-Flash. Placeholder text to be replaced by the corresponding values are in red.

System Prompt for the LLM-Judge during evaluation and **RL** to compute accuracy

Compare the model prediction and the ground truth and determine if they convey the same meaning for the question:

Question: {question}

Model Prediction: {hypothesis}
Ground Truth: {reference}

You MUST respond with the verdict as 'True' if they match semantically or 'False' if they don't match.

Answer in the following format:

```
Reasoning: <Reasoning for the verdict>
Verdict: <True/False>
```

Figure 3. System Prompt for the LLM-Judge during evaluation and **RL** to compute accuracy. Placeholder text to be replaced by the corresponding values are in red.

SAGE Stage-1: Context VLM System Prompt

You are a specialized Context VLM (Video Language Model) designed to analyze video content and determine the appropriate context for further processing. Your primary functions are to:

- Analyze the given video and query
- Recommend the next appropriate tool or sequence of tools
- Suggest specific arguments to pass to those tools

Your output MUST follow this structure and MUST be between the <json> and </json> tags:

```
<json>
{
  "video_context": <visual_context>,
  "query_intent": <user's_intent>,
  "final_answer": "Direct and concise answer to the user's query, if and only if the query is
      answerable based on current context. Otherwise, this should be null.",
  "recommended_tools": {
    "needed": true | false ,
    "why_no_tool": "Only if no more tool call is needed",
    "tool_calls": [
      {
        "rationale": "Why this tool is the best next step",
        "name": <name_of_tool>,
        "arguments": {
          "arg1": <value1>,
          "arg2": <value2>
        }
      }
    ]
  }
}
</json>
```

The available tools are: <<<tools>>>

Figure 4. SAGE Stage-1: Context VLM System Prompt. Placeholder text to be replaced by the corresponding values are in red.

SAGE Stage-2: Iterative Reasoner System Prompt

You are a reasoning agent. Your primary goal is to determine whether the available visual context and tool call information contains sufficient information to answer the user's query. If not, recommend which tools to invoke next, with appropriate arguments.

Do **not** make assumptions beyond the evidence provided. Avoid fabricating facts.

Output Format. You MUST follow this format and MUST be between the <json> and </json> tags:

```
<json>
{
  "answerable": {
    "verdict": true | false,
    "reasoning": "Why the available information is sufficient or not"
  },
  "final_answer": "If the query is answerable, otherwise null.",
  "recommended_tools": {
    "needed": true | false,
    "why_no_tool": "Only if no more tool call is needed",
    "tool_calls": [
      {
        "rationale": "Why this tool is the best next step",
        "name": <name_of_tool>,
        "arguments": {
          "arg1": <value1>,
          "arg2": <value2>
        }
      }
    ]
  }
}
</json>
```

The available tools are: <<<tools>>>

Figure 5. SAGE Stage-2: Iterative Reasoner System Prompt. Placeholder text to be replaced by the corresponding values are in red.

System Prompt for the **ground-event** tool

Given the below event, identify the timestamps for the event in the video.
You are given the snippet belonging to the period between <<<begin>>> and <<<end>>> (in HH:MM:SS format) of the original video.
You should set the start and end timestamps in your answer accordingly to align it to the original video.
If the event does not occur, set start and end to null.

Event:
<<<event>>>

Output Format. You MUST follow this format and MUST be between the <json> and </json> tags:

```
<json>
{
  "name": "the name of the event",
  "timestamps": {
    "start": "start_time", #HH:MM:SS
    "end": "end_time" #HH:MM:SS
  }
}
</json>
```

Figure 6. System Prompt for the **ground-event** tool. Placeholder text to be replaced by the corresponding values are in red.

System Prompt for the reasonable-tool ($s_{\text{reasonable-tool}}$) step reward during **RL**

Below is the reasoning trace for calling a sequence of tools for finding the answer to the question:

Question: {question}

Reasoning Trace: {reasoning_trace}

Predicted Answer: {predicted_answer}

You MUST respond with the verdict as 'True' if the reasoning trace makes sense for the question leading to the predicted answer or 'False' if it doesn't.
You MUST penalize repetitive tool calls if they are not needed.
Answer in the following format:

```
Reasoning: <Reasoning for the verdict>
Verdict: <True/False>
```

Figure 7. System Prompt for the reasonable-tool ($s_{\text{reasonable-tool}}$) step reward during **RL**. Placeholder text to be replaced by the corresponding values are in red.

Prompt for evaluating DIRECT baselines

You will be given a question about a video. You are provided frames from the video, sampled evenly across the video.

Transcript: <<<asr_transcript>>>

Question: <<<question>>>

Respond to the user's question.

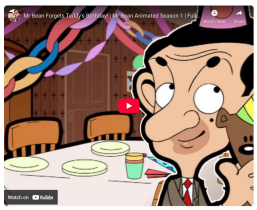
Figure 8. Prompt for evaluating DIRECT baselines. Placeholder text to be replaced by the corresponding values are in red.

YT-ID: EQZ9wtYMfyM



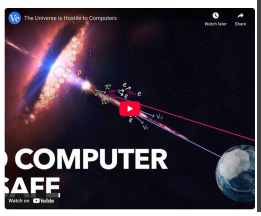
```
{ "index": 15, "type": "open_ended", "difficulty": "medium", "difficulty_rationale":
"Requires remembering the specific people and the lie.", "modality": "both",
"modality_rationale": "Visual: See Joey's gestures. Verbal: Hear him tell Monica what to
say.", "answer": "Joey tries to get Monica to call his parents. He suggests blaming the
Post Office, or 'the Irish'.", "question": "Who does Joey try to get Monica to call
regarding their wedding invitations, and what false reason does he suggest Monica give for
the invitations being late?", "options": null, "requires_web_search": false,
"why_web_search": null, "final_timestamp": "00:27:03", "start_timestamp": "00:25:40",
"end_timestamp": "00:26:55", "compute_percent_video_parsed": "(1615/1623)*100",
"percent_video_parsed": 99.51 }
```


YT-ID: 3HVdiXNa4ak




```
{ "index": 4, "type": "mcq", "difficulty": "easy", "difficulty_rationale": "The action is
clearly shown and is a key plot point.", "modality": "both", "modality_rationale": "Mr.
Bean's actions regarding Teddy's birthday are central to the early plot.", "answer": "He
tries to throw a party and give Teddy a gift.", "question": "How does Mr. Bean react when
he realizes he forgot Teddy's birthday?", "options": [ "He ignores it completely.", "He
tries to throw a party and give Teddy a gift.", "He buys himself a new toy instead.", "He
calls Irma and complains." ], "requires_web_search": false, "why_web_search": null,
"final_timestamp": "00:53:47", "start_timestamp": "01:30:00", "end_timestamp": "02:30:00",
"compute_percent_video_parsed": "((150)/3227)*100 = 4.65", "percent_video_parsed": 4.65 }
```

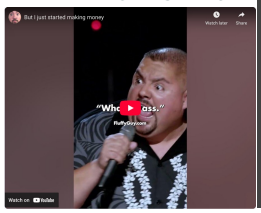

YT-ID: AaZ_RSt0KP8





```
{ "index": 4, "type": "open_ended", "difficulty": "hard", "difficulty_rationale": "The
question requires detailed observation of the animated diagrams and understanding of the
technical explanation provided simultaneously.", "modality": "both", "modality_rationale":
"The process of an SEU is explained through animated diagrams of electron movement in
semiconductors (visual) while also being verbally described in detail.", "answer": "A
Single Event Upset (SEU) occurs when an energetic alpha particle strikes a semiconductor.
It creates electron-hole pairs, causing electrons to accumulate in a well. This can flip a
bit's state (e.g., from a 1 to a 0) even though the device is not physically damaged. It's
a 'soft' error.", "question": "Explain the 'Single Event Upset' (SEU) phenomenon in
semiconductors as described in the video, illustrating how an alpha particle can flip a
bit.", "options": null, "requires_web_search": false, "why_web_search": null,
"final_timestamp": "00:23:01", "start_timestamp": "00:04:08", "end_timestamp": "00:04:46",
"compute_percent_video_parsed": "((286)/1381) * 100", "percent_video_parsed": 20.71 }
```


YT-ID: IPVshfLSmmo




```
{ "index": 12, "type": "open_ended", "difficulty": "hard", "difficulty_rationale": "This
requires identifying the comedian and knowing his well-known stage persona, which is not
explicitly stated in the clip itself.", "modality": "both", "modality_rationale": "The
person is visually present, but their identity and stage name require external knowledge.
The website 'FluffyGuy.com' hints at the persona, connecting visual to external info.",
"answer": "The comedian is Gabriel Iglesias. His stage persona is often referred to as
'Fluffy,' which is also hinted at by the 'FluffyGuy.com' watermark.", "question": "Who is
the comedian performing in this clip, and what is his well-known stage persona?",
"options": null, "requires_web_search": true, "why_web_search": "While 'FluffyGuy.com' is
visible, the name 'Gabriel Iglesias' is not explicitly stated in the dialogue or visually
(except perhaps for 'Gabriel' at 00:00:03, which may not be enough for identification if
unfamiliar with him). A web search would confirm his identity and stage persona.",
"final_timestamp": "00:00:55", "start_timestamp": "00:00:00", "end_timestamp": "00:00:55",
"compute_percent_video_parsed": "(55 / 55) * 100 = 100.00", "percent_video_parsed": 100 }
```

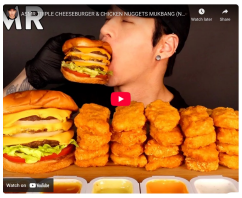

YT-ID: 4ZxNqe-WQDU



```
{ "index": 11, "type": "open_ended", "difficulty": "medium", "difficulty_rationale":
"Requires attention to both commentary and visual cues (on-screen graphics for
positions).", "modality": "both", "modality_rationale": "The question asks for the top
three drivers' names, which are visible on the left-hand side of the screen, and their
associated teams, which are also visible. The commentators confirm their positions.",
"answer": "The top three drivers in qualifying were Max Verstappen (Red Bull Racing RBPT),
Lando Norris (McLaren Mercedes), and Kimi Antonelli (Mercedes).", "question": "Based on
the final qualifying results board, name the top three drivers and their respective
teams.", "options": null, "requires_web_search": false, "why_web_search": null,
"final_timestamp": "00:07:07", "start_timestamp": "00:06:48", "end_timestamp": "00:07:07",
"compute_percent_video_parsed": "((7 + 7 * 60) / (7 * 60 + 7)) * 100 = (427 / 427) * 100 =
100.00%", "percent_video_parsed": 100 }
```


Figure 9. **QnA Pairs Qualitative Samples**. Given our system prompt, Gemini-2.5-Flash can generate high-quality QnA pairs of varying difficulty levels and types (open-ended and MCQ) that cover the entire video.

YT-ID: zqiUYeDMtfI




```
{ "index": 6, "type": "open_ended", "difficulty": "easy", "difficulty_rationale": "The brand name of the beverage is clearly visible on the can as it's being poured.", "modality": "visual", "modality_rationale": "The name of the beverage is written on the can, visible in the video.", "answer": "Sprite", "question": "What brand of soda is poured into the glass?", "options": null, "requires_web_search": false, "why_web_search": null, "final_timestamp": "00:11:17", "start_timestamp": "00:02:24", "end_timestamp": "00:03:09", "compute_percent_video_parsed": "(189/677) * 100 = 27.92%", "percent_video_parsed": 27.92 }
```

```
{ "index": 7, "type": "mcq", "difficulty": "hard", "difficulty_rationale": "The exact 'contains' information is very small print and quickly shown, requiring careful observation.", "modality": "visual", "modality_rationale": "The information about allergens is in very small print on the sauce packets, requiring precise visual attention.", "answer": "Egg, Milk", "question": "What allergens are listed on the McDonald's Ranch sauce packet?", "options": [ "Wheat, Soy", "Milk, Peanuts", "Egg, Milk", "Soy, Nuts", "Gluten, Dairy" ], "requires_web_search": false, "why_web_search": null, "final_timestamp": "00:11:17", "start_timestamp": "02:23:00", "end_timestamp": "02:25:00", "compute_percent_video_parsed": "(2 * 60 + 25) / 677 * 100 = 21.418%", "percent_video_parsed": 21.42}
```


YT-ID: rqQ1X9QjDVg


```
{ "index": 8, "type": "open_ended", "difficulty": "easy", "difficulty_rationale": "The restaurant's name and street address are clearly visible on a sign.", "modality": "visual", "modality_rationale": "The question asks for text visible in the video.", "answer": "The restaurant is named 'Novita' and is located at 102 East 22nd Street.", "question": "What is the name and street address of the Italian restaurant featured with outdoor seating?", "options": null, "requires_web_search": false, "why_web_search": null, "final_timestamp": "00:38:27", "start_timestamp": "05:15:00", "end_timestamp": "05:25:00", "compute_percent_video_parsed": "((5*60*60 + 25*60 + 0)/2307)*100 = 13.04", "percent_video_parsed": 13.04 }
```

```
{ "index": 9, "type": "multiple_choice", "difficulty": "medium", "difficulty_rationale": "Requires careful observation of the specific details on the plaque and distinguishing it from other visual information.", "modality": "visual", "modality_rationale": "The question asks for specific dates and details from a historical plaque, which requires careful visual reading.", "answer": "C. Founded in 1898; Early members include Robert Henri, Frederic Remington, Woodrow Wilson, and Theodore Roosevelt.", "question": "According to the plaque on The National Arts Club building, when was it founded and who were some of its early members?", "options": [ "A. Founded in 1900; Early members include Mark Twain, Thomas Edison, and Andrew Carnegie.", "B. Founded in 1888; Early members include John Singer Sargent, Augustus Saint-Gaudens, and Edith Wharton.", "C. Founded in 1898; Early members include Robert Henri, Frederic Remington, Woodrow Wilson, and Theodore Roosevelt.", "D. Founded in 1920; Early members include F. Scott Fitzgerald, Ernest Hemingway, and Zelda Fitzgerald." ], "requires_web_search": false, "why_web_search": null, "final_timestamp": "00:38:27", "start_timestamp": "06:40:00", "end_timestamp": "07:40:00", "compute_percent_video_parsed": "((6*60*60 + 40*60 + 0)/2307)*100 = 17.34", "percent_video_parsed": 17.34 }
```


YT-ID: pjoQdz0nxf4


```
{ "index": 13, "type": "open_ended", "difficulty": "hard", "difficulty_rationale": "Requires synthesizing information across different parts of the video about multiple extinction events and linking them to volcanic activity.", "modality": "both", "modality_rationale": "The individual mass extinctions are verbally named and visually shown on timelines with associated volcanic regions, which makes it a 'both' modality question.", "answer": "New evidence suggests that at least four of the five biggest mass extinction events (Devonian, Permian, Triassic, and Cretaceous-Paleogene) happened at the same time Earth was furiously spewing gigantic amounts of magma from large igneous provinces like the Kola-Dnieper Provinces, Siberian Traps, Central Atlantic Magmatic Province, and Deccan Traps.", "question": "What new evidence regarding past mass extinction events in Earth's history is presented in the video, linking them to specific geological phenomena?", "options": null, "requires_web_search": false, "why_web_search": null, "final_timestamp": "00:11:39", "start_timestamp": "00:09:00", "end_timestamp": "00:09:17", "compute_percent_video_parsed": "(557 / 699) * 100 = 79.69", "percent_video_parsed": 79.69 }
```


Figure 10. **QnA Pairs Qualitative Samples**. Given our system prompt, Gemini-2.5-Flash can generate high-quality QnA pairs of varying difficulty levels and types (open-ended and MCQ) that cover the entire video.

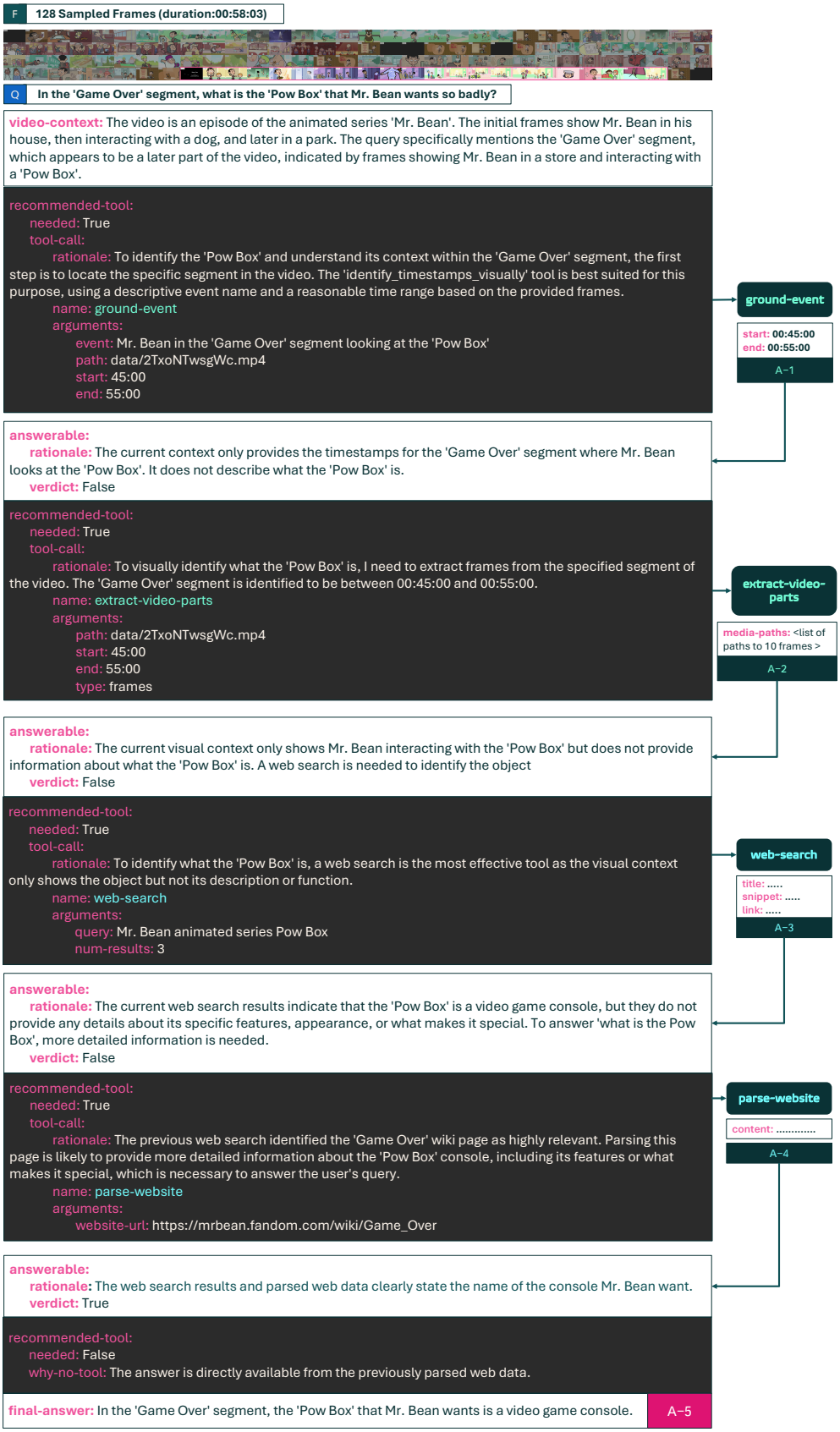



Figure 11. **SAGE Reasoning Trajectory Qualitative Example.** SAGE first coarsely locates the boundaries of the target segment given the sampled frames, followed by a web search to answer the question due to lack of information in the video. The DIRECT Qwen3-VL-8B-Instruct model (which has access to the full video transcript) predicts the following answer: *The "Pow Box" is a fictional device in the "Game Over" segment, likely a humorous or exaggerated gadget that Mr. Bean desires, possibly for its comedic or absurd potential. The exact nature or function of the "Pow Box" isn't specified in the provided transcript, but it's implied to be something desirable or coveted by Mr. Bean, fitting his character's love for quirky, nonsensical objects.*


Figure 12. **SAGE Reasoning Trajectory Qualitative Example.** SAGE accurately transcribes only the target 2-minute segment to answer the user's question. The DIRECT Qwen3-VL-8B-Instruct model (which has access to the full video transcript) predicts the following answer: *(B) 'Thank you, and God bless America.'.*

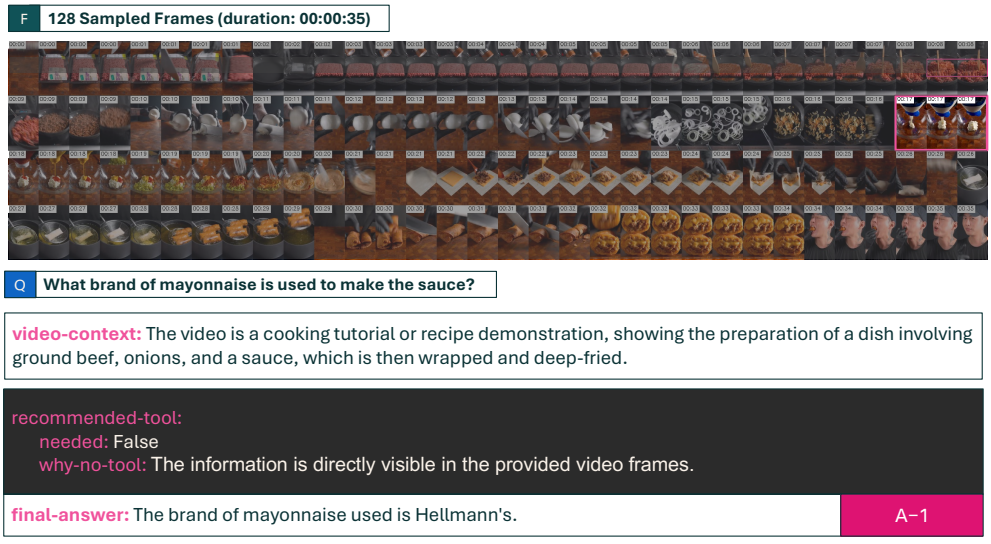



Figure 13. **SAGE Reasoning Trajectory Qualitative Example.** SAGE answers the question in a single turn for the short video. The DIRECT Qwen3-VL-8B-Instruct model (which has access to the full video transcript) predicts the following answer: *Kraft.*