MINERVA: Evaluating Complex Video Reasoning

Supplementary Material

8. Rater Guidelines

All textual data in MINERVA was entirely created by human annotators (raters). All raters are native English speakers with graduate degrees. Here we provide guidelines given to the raters for dataset creation (Sec. 8.1), and for scoring reasoning (Sec. 8.3).

8.1. Dataset Creation

The raters were given the following guidelines before being asked to propose question, answers, decoys and reasoning traces.

8.1.1. Good Questions

- The question should not be easily solvable by looking at just a few frames in the video
- It should not be solvable using only common sense and external knowledge
- It should ask about visual elements in the video (and not just focus on the speech)
- It should not be subjective and should have only one right answer
- It should be complex, and require multiple steps to solve
- It should not be offensive
- Please do not mention any names of humans in the question or the answer unless they are fictional, famous or celebrities
- Each question should belong to at least two question types

8.1.2. Reasoning Traces

- A reasoning step is an action that you would take to break down the question solving process. You can think of them as the building blocks to the solution.
- A good question requires multiple reasoning steps to be performed in sequence to arrive at the answer.
- Without one of the steps, a person should not be able to get the answer.
- Do not add irrelevant information in the steps.
- The final answer can be very short eg. a single word. However, the reasoning steps are the entire process to get to the answer.

8.1.3. Decoys

- A decoy is a wrong answer to the question. We need decoys to create multiple choice questions, like in a multiple choice exam. For our questions, we want to provide 5 options where only one is correct.
- Decoys should be diverse. They should be different enough from each other to not narrow down the scope of the question too much.
- The correct answer should not stand out among the decoys. So, decoys should not have obvious differences to the answer. For example,
  1. Decoys should not be longer or shorter than the true answer.
  2. A decoy should not be an impossible/implausible answer.
  3. A decoy should not make the task easy to solve without watching the video.

8.1.4. Pipelining and Quality Control

The flow is as follows:
Initial Annotation (Curator): A rater famililiar with the subject matter of the video domain (eg for a basketball video, are asked to ensure they know the rules) watches the video and creates 5 distinct questions. This is done concurrently by two raters, to get 10 questions at the end. Each rater works independently to ensure uniqueness and skill diversity.
Peer Review: The initial annotations, including the questions, are then passed to another rater for peer review. This reviewer checks for question complexity and suggests corrections or improvements. (this was implemented during early curation phases and was eliminated later as the raters got more proficient).
Senior Review: Following the peer review, a Senior Reviewer examines the work. The Senior Reviewer makes any necessary final changes to ensure quality and adherence to the established standards.

8.2. Human Study

Goal: You will be given a video, a question, and 5 answer choices. Please watch the video and pick the correct answer. (Like a multiple choice exam). You can watch the video as many times as you like, and you can rewatch various parts of the video. Please take as much time as you require.
Pipelining and Quality Control: We ensure that the raters for dataset creation and the human study are disjoint. 10% of the data is checked by another rater to gauge interrater

agreement for the human study.

8.3. Human Assessments of Reasoning

Goal: You will be given a video and a question. You will also see the outputs of two models which have been asked to answer the question and provide their reasoning.
Your task is to provide scores for the reasoning.
You can judge the reasoning based on the following criteria:
(1) Perceptual correctness: was the relevant information perceived accurately from the video? (eg were the correct objects identified, was the text read properly from the screen, were the relevant events and actions mentioned correctly)
(2) Temporal grounding: were time ranges provided for each piece of information from the video, and if so were they accurate?
(3) Logical reasoning: was the reasoning logically sound, given the information perceived (independent of whether that information was correct)?
(4) Completeness: were any steps skipped in the given answer or left unstated?

For each of the above criteria, please provide a score from 0,1,2.
0 - Doesn't fulfill the criteria at all.
1 - Partially fulfills the criteria.
2 - Completely fulfills the criteria.

9. Model Baselines

9.1. Prompts

```
You will be given a question about a video
and five possible answer options.  You are
provided frames from the video, sampled
evenly across the video.

Transcript:  {asr}
Frames:  {frame1}, ..., {frame N}
Question:  {question}
Possible answer choices:  {answer choices}

Output the final answer in the format
“Final Answer:  (X)” where X is the correct
digit choice.  DO NOT OUTPUT text or any
other words with the full answer.
```

Figure 6. **Direct MCQ prompt for Gemini.**

Section 4.1.1 presents an ablation study on various prompting strategies. Here, we provide details on each strategy. Specifically, Figure 6 illustrates the Direct MCQ prompt, Figure 7 shows the reasoning prompt, and Figure 8 displays the reasoning prompt incorporating the Minerva Rubric.

```
You will be given a question about a video
and five possible answer options.  You are
provided frames from the video, sampled
evenly across the video.

Transcript:  {asr}
Frames:  {frame1}, ..., {frame N}
Question:  {question}
Possible answer choices:  {answer choices}

After explaining your reasoning, output the
final answer in the format “Final Answer:
(X)” where X is the correct digit choice.
Never say “unknown” or “unsure”, or “None”,
instead provide your most likely guess.
```

Figure 7. **Reasoning MCQ prompt for Gemini.**

```
You will be given a question about a video
and five possible answer options.  You are
provided frames from the video, sampled
evenly across the video.

Transcript:  {asr}
Frames:  {frame1}, ..., {frame N}
Question:  {question}
Possible answer choices:  {answer choices}

Provide all steps required to come to
the answer in your reasoning, and the
following rubric will be used to judge
the reasoning:
(1) Perceptual correctness:  was the
relevant information perceived accurately
from the video?
(2) Temporal grounding:  were time ranges
provided for each piece of information
from the video, and if so were they
accurate?
(3) Logical reasoning:  was the reasoning
logically sound, given the information
perceived (independent of whether that
information was correct)?
(4) Completeness:  were any steps skipped
in the given answer or left unstated?

After explaining your reasoning, output the
final answer in the format “Final Answer:
(X)” where X is the correct digit choice.
Never say “unknown” or “unsure”, or “None”,
instead provide your most likely guess.
```

Figure 8. **Reasoning MCQ prompt for Gemini with the Minerva Rubric.**

9.2. Implementation Details

9.2.1. Hyperparameters

Hyperparameters for all our models are provided in Table 5.

~~~
You are an expert at grading student answers to questions about videos.  For each video, you will get a question about the video, the correct reasoning, and
the final answer.  You will then get the reasoning from the student, and a set of criteria.  Given this criteria, please provide a score from 0, 1 or 2 for each
criterion that will assess the student’s work.


**Criteria:**
(1) Perceptual correctness: was the relevant information perceived accurately from the video?
(2) Temporal grounding: were time ranges provided for each piece of information from the video,
and if so were they accurate?
(3) Logical reasoning: was the reasoning logically sound, given the information perceived (independent
of whether that information was correct)?
(4) Completeness: were any steps skipped in the given answer or left unstated?

For each of the above criteria, please provide a score from 0,1,2.
0 - Doesn’t fulfill the criteria at all
1 - Partially fulfills the criteria
2 - Completely fulfills the criteria
Please produce the score in the JSON format:
```
{"Perceptual correctness": <score that is 0,1,2>, "Temporal grounding": <score that is 0,1,2>, "Logical reasoning": <score that is 0,1,2>,
"Completeness": <score that is 0,1,2>}
```

**Examples:**

**Question:** What has to happen, according to the performer, for the wishing audience member’s wish
to come true? The wishing audience member must stand up., The wishing audience member must come up
onstage., The entire audience must close their eyes., The entire audience must imagine the wishing
audience member naked., The wishing audience member must receive truth.,
**Reference Answer:** I watched the sequence in the video where the performer dresses up at the
"Magic Magic Wish Man" from 03:18 to 04:00, and singles out an audience member to make a wish at
03:38. I then listened for the performer to tell the audience how to make the wish come true, and
from 03:50 to 03:55, he tells the entire audience that they must close their eyes.
**Student Reasoning** The entire audience must close their eyes.

Output should be:
```
{"Perceptual correctness": 2, "Temporal grounding": 0, "Logical reasoning": 2, "Completeness": 0}
```

**Question:** How many hearts are visible in the picture of the finished building at 00:37?8., 10.,
7., 11., 9.,
**Reference Answer:** I watched the video until the indicated time code of 00:37. From there, I
counted the number of red and blue hearts visible in the photo: 4 on the columns, 1 on the entrance
roof, 1 on the left edge of the building, and 3 on the top facade. This comes to a total of 9 visible
hearts in the photo.
**Student Reasoning** There are five hearts visible on the building in the image at 00:37.

Output should be:
```
{"Perceptual correctness": 1, "Temporal grounding": 2, "Logical reasoning": 0, "Completeness": 1}
```

**Question:** How did the woman in green react to the woman in blue crying? She became worried., She cried with her., She became annoyed., She became indifferent.,
She showed empathy.,
**Reference Answer:** I watched the video until 1:58, where the woman in blue cried. Afterward, the woman in green asked her if she plays video games. The woman in
blue responded by saying she plays SIMS and pretends that they’re her family. Then, the woman in green looked at the man in red and widened her eyes, a clear sign
that she was worried.
**Student Reasoning** At around 2 minutes into the video, there’s a scene where the woman in blue starts crying while sitting on the couch. The woman in green, who
is seated next to her, leans forward, places a hand on her shoulder, and speaks to her in a comforting manner. This action clearly indicates that she is showing
empathy towards the woman in blue. There is no indication of worry, annoyance, indifference, or crying along, which rules out options A, B, C, and D.

Output should be:
```
{"Perceptual correctness": 0, "Temporal grounding": 1, "Logical reasoning": 1, "Completeness": 0}
```

**Question:** How many passes did the white team make right before they scored their first goal? 4., 1., 2., 5., 3.,
**Reference Answer:** I searched for the white team’s first score, which happened at 00:42. I watched the clip that showed the entire play, which is from 00:31 to
00:43. I counted one pass at 00:39 and one pass at 00:41. The player who receives the pass at 00:41 shoots into the goal. To confirm, I watched the replay from
00:44 to 00:47. At 00:46, the puck bounces off a white player’s stick and into the goal. This makes the shot taken at 00:41 a pass and assist instead of a proper
shot, which means the white team made 3 passes before they scored their first goal.
**Student Reasoning** In the frames leading up to the first goal scored by the team in white, we can see a sequence where the puck is passed between players. The
sequence goes as follows:\n\n1. A white team player makes a pass to a teammate in the offensive zone.\n2. The teammate receives the puck and makes a quick pass to
another player.\n3. The third player receives the pass and immediately shoots on goal.\n\nThis sequence involves a total of three passes made by the white team
right before the goal is scored.\n\nFinal Answer: (E)

Output should be:
```
{"Perceptual correctness": 2, "Temporal grounding": 0, "Logical reasoning": 2, "Completeness": 2}
```

**Re-iterating the Criteria**
(1) Perceptual correctness: was the relevant information perceived accurately from the video? (eg were the correct objects identified, was the text read properly
from the screen, were the relevant events and actions mentioned correctly)
(2) Temporal grounding: were time ranges provided for each piece of information from the video, and if so were they accurate?
(3) Logical reasoning: was the reasoning logically sound, given the information perceived (independent of whether that information was correct)?
(4) Completeness: were any steps skipped in the given answer or left unstated?

**Final Instruction and Input**
Please now produce the scores (0, 1, or 2) in the correct JSON format for the following question, reference answer, and student reasoning, following the criteria.
**Question:** {question} {formatted_answer_options}

**Reference Answer:** {reference_reasoning} Final Answer: {reference_final_answer}

**Student Reasoning:** {model_reasoning}

JSON output and justification:
~~~

Figure 9. **Scoring prompt for MiRA analysis.** We show the main “reference-based” prompt above, which includes reference examples with ground-truth reference reasoning traces, as well as the reference ground-truth reasoning trace for the question + reasoning being evaluated. The “reference-free” version run for comparison in the main paper omits this. In both prompts we provide our MINERVA rubric.

Table 5. **Hyperparameters for all model baselines**

Method	# of Frames	ASR	Hyperparameters (seeds, temperature, etc)

InternVideo2.5 [46]	256	✓	image size=448,
			temperature=0,
			top-p=0.1 (default),
			beams=1,
			sample=False,

Qwen2.5-VL-72B [6]	768*	✓	frames=2fps up to 768 frames (default),
			seed=default,
			sampling=default

VideoLLaMA3-7B [51]	180*	✓	frames=1fps up to 180 frames (default),
			seed=default,
			sampling=default

Deepseek-R1:32b	0 (blind)	✓	seed=default,
			temperature=1 (default),
			top-p=default,

GPT-4o [1]	250	✓	version=gpt-4o-2024-08-06,
			seed=default,
			top-p=default,
			temperature=1 (default),
			image resolution model=low

GPT-4.1 [2]	256	✓	version=gpt-4o-2024-08-06,
			seed=default,
			top-p=default,
			temperature=1 (default),
			image resolution model=low

Claude 3.5 Sonnet v2 [4]	64	✓	image_size=448, other parameters = default

Gemini 1.5 Pro [39]	all	✓	temperature=0,
			seed=default,
			sampling=default

Gemini 2.0 Flash [39]	all	✓	temperature=0,
			seed=default,
			sampling=default

Gemini 2.5 Flash Thinking [43]	all	✓	temperature=0,
			seed=default,
			sampling=default

Gemini 2.5 Pro Thinking [43]	all	✓	temperature=0,
			seed=default,
			sampling=default

OpenAI o1 [23]	64	✓	image_size=448,
			reasoning_effort=medium,
			other parameters = default

9.2.2. API Access Dates

We accessed each API for all API-based models between Feb 24 and Mar 7, 2025.

10. Ablations

We show a fine-grained frame ablation for Gemini 2.0 Flash in Fig. 10. Results appear to saturate around 256 frames for this model, which also sets the state-of-the-art on MINERVA.

11. Statistics for MINERVA

We provide a distribution of skill types for the questions in the dataset in Fig. 11. Note each question in the dataset requires two or more skills, and hence we show an upset plot of combinations of skills. We omit Object Recognition and Temporal Reasoning, as they are required for almost all questions.

12. Reasoning Analysis

Examples of the MINERVA taxonomy are provided in Table. 8. For scoring using an LLM (MiRA), we examine the correlation between LLM and human judgement. We first conduct an experiment to determine which LLM to use (Tab. 6). In general, human correlation scores increase with LLM size. We also experiment with an MLLM conditioned on video (1 fps, Gemini 2.0), and find that performance decreases slightly (row 5 vs. 3), likely due to context dilution from frame tokens. This highlights the advantage of our ground-truth reasoning: it provides a condensed form of the relevant information, enabling a cheaper, text-only, metric. In general, for text-only Gemini 2.0, correlation with human judgments are reasonable (> 0.3 Pearson r scores for all 4 axes), and are specially high for T, P and C rubric criteria. We therefore use this model as our metric in this work.

The scoring prompts used for the MiRA analysis are in Figure 9. Results showing MiRA scores for all 4 axes of the rubric on the full dataset can be found in Fig. 12 and Table 7. Table 9 shows how the quality of reasoning traces can differ dramatically even though the models achieved comparable MCQ performance.

Table 6. **LLM-as-a-judge ablation for scoring reasoning traces.** We perform an analysis of reasoning traces with our proposed MINERVA Rubric, with different LLM-based methods, and show correlation with human judgment (Pearson r). We observe that larger models have higher human correlations. In general, for Gemini 2.0, correlation with human judgements are reasonable, specially for T, P and C rubric criteria.

Method	**Modality**	**Rubric Criteria: Pearson *r***				**Model Size**
		T	P	L	C	
Gemma-4B	T	0.46	0.35	0.32	0.31	4B
Gemma-27B	T	0.77	0.46	0.23	0.59	27B
Gemini 2.0	T	0.80	0.58	0.41	0.58	Unknown, >27B
GPT-4.1	T	0.82	0.59	0.48	0.57	Unknown, >27B
Gemini 2.0	(MLLM) T+V	0.77	0.50	0.38	0.53	Unknown, >27B

13. Ablations

Thinking ablations for Gemini 2.5 Pro are provided in Table 12. Prompting Ablations are provided in Table 13. We find that asking the model to perform step-by-step reasoning rather than directly producing an answer results in a significant boost to MCQ accuracy. What is interesting however, is that explicitly providing the rubric in the prompt improves the final score even further (the reasoning outputs also improve, as shown by an automatic LLM judge (MiRA) which is described in Sec. 5.1.1). Note that this improvement comes with minimal extra inference-time compute (no multiple calls needed), and our rubric was designed to be as general as possible (does not contain any few-shot examples specific to the dataset). This suggests that asking models to provide reasoning along the four axes we identified in the rubric for video can actually improve final outcomes as well.

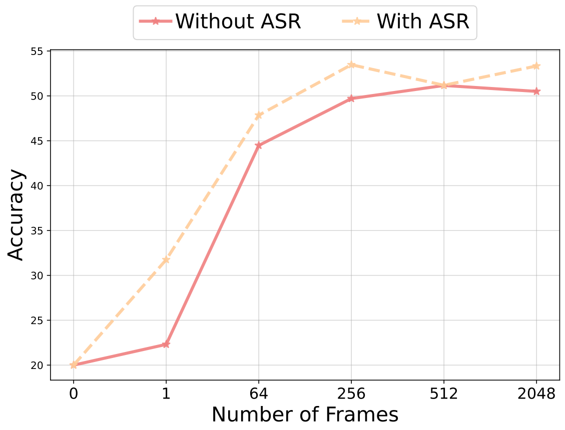

Figure 10. **Results with varying number of frames using Gemini 2 Flash:** Results appear to saturate around 256 frames (note: x-axis is not linear scale).

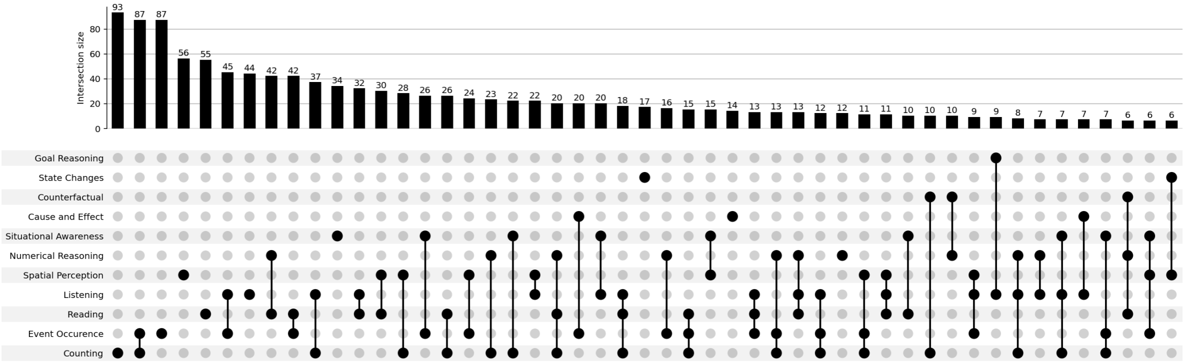

Figure 11. **Distribution of skill types:** Given each question requires two or more skills, we show an upset plot of combinations of skills. We omit Object Recognition and Temporal Reasoning, as they are required for almost all questions.

Table 7. **Benchmarking performance on MINERVA.** We report multiple choice accuracy (MCQ-Acc.) and MiRAscores normalized to be between 0 and 1. P: Perceptual Correctness, T: Temporal Localization: L: Logical Reasoning: C: Correctness * indicates FPS sampling up to frame limit, following optimal settings from prior work[6, 51].

Method	**# of Frames**	**ASR**	**MCQ-Acc. %**	**MiRA**				
				P	T	L	C	Total
Random	-	-	20.00					
Open-source								
Qwen2.5-VL [6]	768*	✓	35.05	0.42	0.39	0.80	0.68	0.57
VideoLLaMA3 [51]	180*	✓	35.91	0.40	0.10	0.72	0.31	0.39
InternVideo2.5	256	✓	35.18	0.46	0.16	0.77	0.34	0.43
Proprietary								
Claude3.5 Sonnet v2 [4]	64	✓	31.28	0.47	0.55	0.78	0.73	0.63
OpenAI o1 [23]	64	✓	43.48	0.52	0.52	**0.86**	**0.88**	0.69
GPT-4o [1]	250	✓	45.54	0.57	0.67	0.77	0.79	0.70
Gemini 1.5 Flash [39]	2048	✓	49.31	0.57	0.68	0.82	0.81	0.72
Gemini 2.0 Flash [39]	256	✓	**53.47**	**0.62**	**0.75**	0.83	0.82	**0.75**

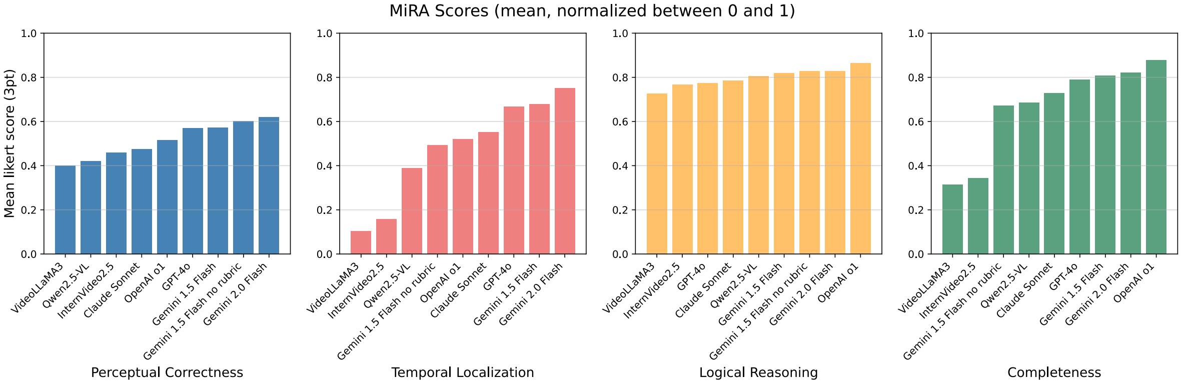

Figure 12. **Analysis of model produced reasoning traces on MINERVA with MiRA:** We report MiRA scores on reasoning traces for all 4 different axes in the MINERVA rubric. Models are presented in ascending order of scores. Results are lower for perceptual correctness and temporal grounding than logical reasoning and completeness. This suggests that when asking frontier multimodal models (which contain heavy, well-trained LLMs) reasoning-heavy video questions, they tend to provide plausible sounding detailed reasoning traces, however they lack key perceptual insights and temporal grounding in the video.

Table 8. **Taxonomy of Video Reasoning Errors.** Examples are provided from MINERVA, and the model used is Qwen2.5-VL-72B. Note how in the first example, the model fails to correctly perceive the score to be 62-60 at the end, which leads to the wrong conclusion, while in the second example the model incorrectly localizes the announcement to start at 03:04, instead of 02:15. In the fourth example, the model concludes that the man in the cage is called Ryan simply because another character uses the name Ryan, which is not correct. In the final example, the model simply produces the final answer without showing any steps.

Error Category	Question	Answer	Ground truth Reasoning	Model Reasoning
Temporal Localization	What happens after the announcer announces the reward for winning the match?	The crowd loudly boos.	At 02:15 - 02:21, I heard the announcer announce "Whoever wins out of you two enters the rumble last". Immediately after, I then heard the surrounding crowd loudly boo. Therefore, the crowd booed after the announcer announced the reward for winning the match.	The announcer explains that the winner of the match will enter the rumble last, and the loser will enter first. Immediately following this announcement (around 03:04-03:06), the video shows the two wrestlers beginning to fight.
Perceptual Correctness	What is the significance of Aja Wilson's last shot attempt?	The game would have been tied if it went in.	At 00:30-00:40, I heard the announcer say that Aja Wilson was dominant all game for the Gamecocks as I saw a player make a shot. At 00:37, the player turned and walked away and I saw the number "22" and the word "Wilson" written on her back, indicating that Aja Wilson wears the number 22. Then, I watched this player's last shot attempt from 01:47-01:52. The shot was taken close to the basket and inside the 3-point line, so it is a 2-point attempt. At 01:53, I saw the final score, which would've been 62-62 if Aja Wilson had made her last shot. Therefore, the game would have been tied if the shot went in.	The video shows that the game was tied at 60-60 when Aja Wilson attempted her last shot. The score shown on the screen at 1:23 is 60-60. If Wilson's shot had been successful, South Carolina would have taken the lead and likely won the game, as there was little time remaining.
Logical Reasoning	What is the mans name that is sitting down inside of the cage?	Jacob.	I watched as the two characters from the beginning of the video enter the doorway. At 00:39, one of the characters yells "Jacob" as he runs towards the cage. Inside the cage we can see a man sitting, at 00:41, who responds to the name "Jacob".	The man inside the cage is referred to by the name ’Ryan’ when another character says, ’Hey! Hey, it's Ryan! Jesus, what happened to you..?’ This dialogue occurs around the 45-second mark of the video.
Completeness	If each settlement is surrounded by 3 tiles with numbers, what is the sum of the numbers on the tiles surrounding both the blue and black settlements at 05:10?	47.	I found the blue and black settlements on the board at 05:16. At the same time, I noticed the blue settlement was surrounded by 3 tiles with the numbers "5", "8", and "10". Likewise, I saw the black settlement was surrounded by 3 tiles with the numbers "9", "11", and "4". I then calculated 5+8+10+9+11+4=47. Therefore, the sum of the numbers on the tiles surrounding both the blue and black settlements at 05:10 is 47.	The sum of the tiles gives a total of 49. Therefore, the correct answer is 49.

Table 9. **Qualitative Comparison of Model Reasoning:** We compare reasoning traces from InternVideo and Qwen2.5 for questions where both models got the correct multiple choice answer. InternVideo2.5 generally provides shorter reasoning traces with less evidence from the video (first to third examples) or simply restating the question and answer (second example). It also hallucinates (fourth example) or simply guesses when it fails to extract the evidence from the video (fifth example).

Question	**Answer**	**Ground truth Reasoning**	**Qwen2.5-VL Reasoning**	**InternVideo2.5 Reasoning**
What is the unique advantage Player 32 from Bristol Central has over the opposing team?	He is the tallest player among both teams.	I looked at the video in its entirety to locate instances where Player 32 from Bristol Central scored points or blocked the opponents. At 00:21 he makes his first block, at 00:33 he scores his first point, at 00:54 and 01:01 he makes 2 more blocks, and at 01:30 he makes his first dunk. From looking at high angles shots of the basketball court, such as the one at 01:08, it is evident that Player 32 is the tallest player among both teams. This information helped me to conclude that Player 32's unique advantage is his height.	Throughout the video, Player 32 from Bristol Central is consistently seen dominating the paint area due to his height advantage. This is evident when he successfully grabs rebounds and scores over the heads of the opposing players. The visual evidence shows that he towers over other players during key moments like the 1:43 mark where he secures a rebound despite being surrounded by defenders. His height allows him to reach higher for shots and blocks, which is a significant advantage in basketball.	Player 32 from Bristol Central is consistently shown as the tallest player on the court, making him a dominant presence in the paint. This advantage allows him to score more points and outplay the opposing team's players.
What does the basketball coach do after the interviewer asks his first question?	He scratches his head.	I looked for the interviewer's first question. This happens at the beginning of the video, from 00:00, and it goes to 00:04. At this point, the interviewer finishes his question and stops talking. At this point, I watch the basketball coach to see his response. As he responds, he raises a hand to the side of his head and scratches.	After the interviewer asks the first question, the basketball coach can be seen scratching his head while responding. This gesture occurs shortly after the initial question is posed, indicating a moment of reflection or emphasis as he begins to elaborate on the team's performance. The visual evidence from the video clearly shows this action, aligning with option (C).	After the interviewer asks his first question, the basketball coach is seen scratching his head.
How many shots on goal were taken that were saved by the goalie of either team, not including replays?	6.	I watched the video for all shots taken on a goal. After discarding the shots that scored, the video shows ten shots taken that were saved: 00:16, 00:17, 00:23, 00:25, 00:27, 01:01, 01:37, 01:40, 01:45, 02:05. I noticed that four of these were replays: 00:17, 00:27, 01:40, 01:45. After subtracting replays from shots taken and saved, the answer is (1) Six.	Based on the video and the provided transcript, we can identify several instances where shots were taken and saved by the goalies. At around 00:25, a shot was deflected off the crossbar, indicating a save. Around 00:47, another shot was mentioned but not scored, suggesting a save. At 01:15, a shot was described as being saved. Additionally, at 01:58, a shot was blocked by a stick. Finally, at 02:09, a shot was let go but not scored. These five saves, along with one more visible in the video near the end of the first period, total six saves. Therefore, the correct answer is E.	The video shows several shots on goal that were saved by the goalie of either team. By counting these shots, we can determine the answer.
What two cards, including their suits, are Hawk Tuah holding as part of her 2-pair hand in hand three?	King of Hearts, Queen of Spades.	I listened to the narrator introduce hand 3 at 04:28. Just after that time, **at 04:30**, I watched as Hawk Tuah's hand was displayed on the left side of the screen, to show a King of Hearts and Queen of Spades just above her name.	In the video, during the discussion of Hand Three, the on-screen graphics clearly show that Hawk Tuah is holding the King of Hearts and Queen of Spades. This can be seen around the timestamp where the flop is revealed and the players' hole cards are displayed. The visual confirmation matches option (C), making it the correct choice.	In the video, Hawk Tuah is shown holding a 2-pair hand in hand three. The cards she is holding are King of Hearts and Queen of Spades. This information is visible in the video at around 20 to 25 seconds.
After the winner of the final hand stands up, where is the next person who stands up located in relation to the winner?	To their left.	I watched the winner of the final poker hand, Vegas Matt, standing at 07:18 while the AI narrator says that he won the hand. I then watched the video for the next person to stand up, which I saw at 07:20 as a woman stood and leaned forward over the table to gather poker chips. **I noted this woman's location in relation to Vegas Matt - she was directly to his left at the table.**	In the video, after the final hand concludes and the winner, Vegas Matt, stands up, the next person to stand up is located to his left. This can be observed around the timestamp near the end of the video where the final hand is played. The seating arrangement shows that the person to the left of the winner is the next to stand, likely due to the elimination process in the tournament. The other options such as 'across the table', 'directly behind', 'to their right', or 'directly across the table' do not match the visual sequence of events captured in the video.	The video does not provide a clear view of the relative positions of the people standing up after the winner of the final hand. However, based on the information provided in the video, it is possible to infer that the next person who stands up is located to the left of the winner.

Table 10. MCQ scores by video category (see Fig. 3, middle)

	Qwen2.5-VL	VideoLLaMA3	InternVideo2.5	Claude 3.5 Sonnet v2	OpenAI o1 (64f)	GPT-4.1	Gemini 2.5 Pro (Thinking)	Humans
Misc Sports	29.76	34.52	39.29	29.76	39.29	50.00	57.14	95.24
Tennis	27.87	36.07	27.87	32.79	44.26	36.07	55.74	95.08
Travel	30.77	23.08	25.96	27.88	35.58	45.19	65.38	91.35
Motorsports	41.67	35.42	37.50	43.75	56.25	68.75	72.92	95.83
Tech/AI	31.58	31.58	28.95	34.21	42.11	50.00	63.16	76.32
Maths	39.53	25.58	25.58	34.88	44.19	48.84	48.84	90.70
Short Films	48.14	48.55	47.11	44.01	54.13	68.60	76.65	97.31
Basketball	40.32	32.26	41.94	37.10	56.45	61.29	58.06	95.16
Animals	19.01	22.31	23.14	21.49	28.10	41.32	47.11	88.43
Board Games	34.04	31.91	29.79	26.24	39.72	44.68	65.96	95.74
Physics	16.33	30.61	32.65	24.49	34.69	38.78	61.22	81.63
Chess	21.82	32.73	27.27	33.64	35.45	47.27	56.36	92.73
Cooking	30.68	29.55	29.55	20.45	25.00	44.32	60.23	85.23
How-To	31.71	21.95	32.93	29.27	31.71	50.00	59.76	80.49

Table 11. MCQ scores by skill (see Fig. 3, left)

	Qwen2.5-VL	VideoLLaMA3	InternVideo2.5	Claude 3.5 Sonnet v2	OpenAI o1 (64f)	GPT-4.1	Gemini 2.5 Pro (Thinking)	Humans
Temporal Reasoning	32.88	35.16	33.03	32.27	40.33	51.29	61.19	93.15
Counterfactual	29.70	21.78	25.74	37.62	32.67	44.55	55.45	88.12
Spatial Perception	33.90	37.33	39.38	31.85	44.86	55.14	63.36	94.86
Listening	36.50	34.75	31.75	38.00	46.00	55.50	67.50	92.00
State Changes	34.43	34.43	36.07	32.79	42.62	57.38	59.02	88.52
Event Occurence	33.40	34.14	35.44	33.21	39.52	51.76	62.71	93.69
Cause and Effect	44.57	40.22	33.70	31.52	46.74	66.30	69.57	93.48
Situational Awareness	39.76	38.96	37.35	38.55	43.78	57.03	62.65	95.58
Goal Reasoning	45.00	40.00	42.50	32.50	50.00	62.50	77.50	92.50
Numerical Reasoning	26.82	27.20	24.14	32.57	38.31	41.38	59.39	87.36
Counting	24.07	25.19	28.89	27.78	30.37	40.00	51.67	86.11
Reading	39.23	34.92	33.56	35.37	47.17	56.46	70.98	92.52
Object Recognition	34.20	36.30	36.67	32.22	42.47	53.33	65.43	90.99

Table 12. **Thinking ablations on MINERVA with Gemini 2.5 Pro [43]. MCQ results are provided as Acc.%.**

Model	# Frames	Thinking Off	Thinking On
2.5 Pro	256	63.1	64.7
2.5 Pro	512	62.3	66.0
2.5 Pro	1024	63.9	**66.2**

Table 13. **Prompting Ablations on MINERVA.** Results of Gemini 2.0 Flash, 256 frames + ASR. We provide both MCQ accuracy on the final answers and MiRA on the reasoning traces. † Very few reasoning outputs (by design) to assess.

Prompting Method	MCQ Accuracy	MiRA
Direct Answer	46.47	†
+ Reasoning	51.22	0.65
+ Minerva Rubric	53.47	0.75