

Minerva-Ego: Spatiotemporal Hints for Egocentric Video Understanding

Supplementary Material

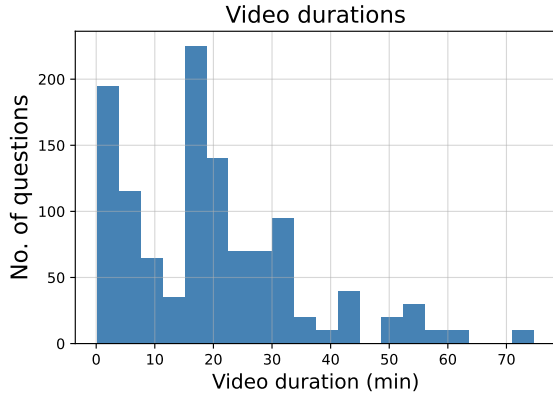


Figure 1. Video Lengths in Minerva-Ego.

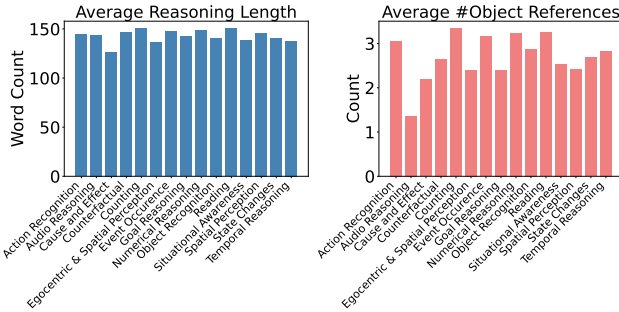


Figure 2. Statistics of reasoning trace lengths and object references broken down by specific question types.

1. Minerva-Ego

1.1. Additional Statistics

The shortest videos is 10 seconds while the longest video is 75 minutes. The mean video length is 20 minutes. The full distribution can be found in Fig. 1.

Figure 2 shows statistics of reasoning lengths and object occurrences by question type. We note that the distribution is flat, other than *Audio Reasoning*, which has fewer object references (as expected).

1.2. Dataset Post-processing details

As described in Sec. ??, our pipeline consists of selection, annotation + verification, and post-processing. We elaborate here on the verification and post-processing stages. We follow best practices and guidelines that have been established in prior work [? ? ? ?], which note that benchmarks often have issues regarding modality biases (e.g. questions that are answerable with only text), and propose solutions

to help mitigate these. The verification stage is described in the main paper, and was conducted through manual human studies, separated reviewer pools, and several rounds of intervention and refinement. In this context, the post-processing stage is a further step of refinement on top of our initial (model-free) process. Here, we use the same model ensemble of open and closed-source text-only language models in [?] (e.g. Deepseek, GPT-4o, Gemini Flash, and Qwen) and take their *consensus* agreement as a set of candidate questions that merit further review. As described in prior work, this technique is designed in order to avoid discarding difficult questions that models may have answered correctly by chance. Notably however, rather than completely discarding all questions in the consensus set like in prior work, this set was subject to one more round of review – if the question could be “corrected” (e.g. the text bias was readily clear and it was clear how a change in the question or answer choices could remove this bias) then this measure was taken (we found that this was the case in the majority of the questions, where “too much” was given away in the question itself). If the question required more than a simple change, the question was removed.

2. Rater Guidelines

All textual data in Minerva-Ego was manually annotated by human annotators (raters). All raters are native English speakers with graduate degrees.

2.1. Annotation Guidelines

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

You will be given a video. For each video, we would like you to propose complex multi-step questions. Questions should be accompanied by 5 answer candidates, one correct answer, and a detailed reasoning trace. The reasoning trace should refer to objects present in the video (more details are provided below).

Questions should require 2 or more of the skills provided below:

1. Temporal Reasoning
2. Counting
3. Cause and Effect
4. Goal Reasoning
5. Egocentric and Spatial Perception
6. Event Occurrence
7. State Changes
8. Reading (OCR)

9. Numerical Reasoning (all math operations other than counting)
10. Object Recognition
11. Action Recognition
12. Counterfactual (“what if”)
 - Important things to keep in mind:
 1. Questions should be multi-step
 2. They should be difficult to solve
 3. Ideally they should involve looking at multiple different time segments of the video
 4. Each question should be cover multiple skills and have multiple reasoning steps
 5. The questions should be phrased with the word “I”, as if the cameraperson is asking the question.
 6. The reasoning traces should use the word “I” to refer to you (the rater, as you are doing the reasoning, and use the word “user” to refer to the cameraperson). Note in the question, I refers to the camera person.
 7. The reasoning traces should map to the “objects of interest”.
 8. Every time that object is referenced (even if it is in a different time step of the video), please use the same annotation.
 9. Please number the “steps” in each reasoning trace, as shown in the examples.
10. All reasoning traces must have at least one time stamp, and at least one object of interest reference.

2.2. Human Study

We used a disjoint pool of 12 raters for the human study, by asking the rater leads to ensure that the same rater who proposed the question (or even saw the video before) is not the same as the one who performs the human study. Hence no rater performing the humans study has seen the video before. All raters had access to the full videos. Rater guidelines are below:

Goal: You will be given a video, a question, and 5 answer choices. Please watch the video and pick the correct answer (A, B, C, D, E). (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.

3. Model Implementation Details

Hyperparameters for all our models are provided in Table 1.

4. Results

4.1. MiRA Reasoning Evaluation

MiRA is a model-based metric that uses Gemini 2.5 Pro for scoring. The exact prompt used to compute the MiRA score appears in Figure 3.

Since using the same LLM for hinting and as a judge could *potentially* bias the results, we repeat the MiRA analysis from Fig. ?? (left) using GPT-5 as a judge, and report the scores in Figure 4. Comparing the two, we can see that the *relative ordering* of the models is preserved. Since this evaluation compares the ground truth reasoning trace with the model’s using a fixed rubric we believe there is less dependence on the judge itself.

It’s worth noting that Minerva-Ego includes multiple-choice answers for every question, enabling MCQ accuracy to serve as a primary, model-free metric.

4.2. Results by Skill/Length

We break down the performance of SotA models by video length in 5. Performance of most models is remarkably robust with increasing context length, but generally declines for videos longer than one hour. 6 breaks down performance by skills. Note that we excluded the following skills because they had less than 50 examples in the dataset: Reading, Cause and Effect, Goal Reasoning, Audio Reasoning and Spatial Perception. Generally, models perform best at questions involving State Changes, Counterfactuals and Situational Awareness, but struggle with Event Occurrences, Numerical Reasoning and Counting.

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?

****Question:**** {question}

****Reference Reasoning:**** {reference_reasoning}

****Reference Answer:**** {reference_final_answer}

****Student Reasoning:**** {model_reasoning}

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 format [Criteria: score]

Examples:

****Question:**** How many times heavier is the second density cube than the first? 2.64., 3.05., 2.86., 1.35., 1.77.,

****Reference Answer:**** I began watching the video and noticed many duplicate items. At 01:57, the narrator says "density cubes," and then shows the cubes. At 02:01, they are all labeled. There are a total of 12 density cubes shown. At 02:07, the narrator takes 2 away to weigh them. The first cube displays a number of 44 on the scale, shown at 02:11. At 02:13, the second cube is weighed. The number 116 appears on the scale. These two numbers are not said out loud. If we divide 116 by 44 and round to 2 decimal places, we get 2.64. Final Answer: 2.64.

****Student Reasoning****

The question asks how many times heavier the second density cube is than the first.

From [01:58] to [02:04], the video shows 12 density cubes.

From [02:00] to [02:04], the video labels the first cube as "acetate" and the second cube as "polypropylene".

From [02:11] to [02:13], the video shows the presenter putting the acetate cube on a scale, which reads 44 grams.

From [02:11] to [02:13], the video shows the presenter putting the polypropylene cube on a scale, which reads 24 grams.

To find how many times heavier the second cube is than the first, we divide the weight of the second cube by the weight of the first cube: $44/24 = 1.83$.

The closest answer choice to 1.83 is 1.77.

Output should be:

- (1) Perceptual correctness: 1
- (2) Temporal grounding: 2
- (3) Logical reasoning: 1
- (4) Completeness: 2

****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. Final Answer: 3.

****Student Reasoning****

The video shows the white team making a series of passes before scoring their first goal. The sequence starts with a player receiving the puck near the blue line, then passing it to a teammate who skates towards the goal. This teammate then makes another pass to a third player who is positioned closer to the net. Finally, this third player takes a shot and scores. This sequence involves three distinct passes, which aligns with option (E). The timestamps and visual cues in the video confirm these actions, showing clear transitions between players handling the puck.

Output should be:

- (1) Perceptual correctness: 2
- (2) Temporal grounding: 0
- (3) Logical reasoning: 2
- (4) 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?

Figure 3. The MiRA rubric prompt, as provided to Gemini (or GPT), used to score the reasoning traces produced by the VLMs being evaluated, as described in Section ??.

Table 1. Hyperparameters for all model baselines

Method	# of Frames	Hyperparameters (seeds, temperature, etc)
Qwen-3 [?]	all @ 1fps	version=Qwen/Qwen3-VL-235B-A22B-Thinking, max-sequence-length=128k
GPT-5 [?]	64	version=gpt-5-2025-08-07, seed=default, top-p=default, temperature=default, image resolution=auto, reasoning effort=medium
GPT-4.1 [?]	64	version=gpt-4.1-2025-04-14, seed=default, top-p=default, temperature=default, image resolution=auto
Claude Sonnet 4 [?]	64	version=claude-sonnet-4-20250514 seed=default, top-p=default, temperature=1.0, thinking budget=10,000
Gemini 2.5 Flash Thinking [?]	64	temperature=default, seed=default, sampling=default
Gemini 2.5 Pro Thinking [?]	64	temperature=default, seed=default, sampling=default

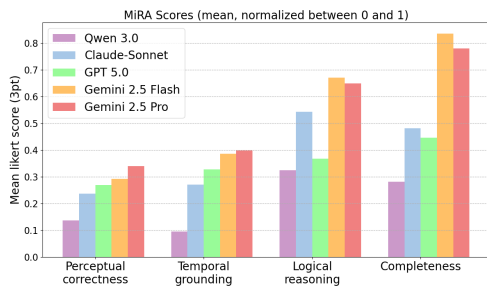


Figure 4. MiRA scores of the model reasoning traces using GPT-5 as a judge.

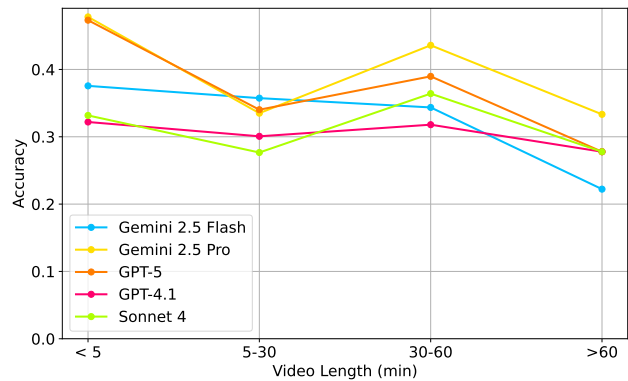


Figure 5. Accuracy for videos of different lengths

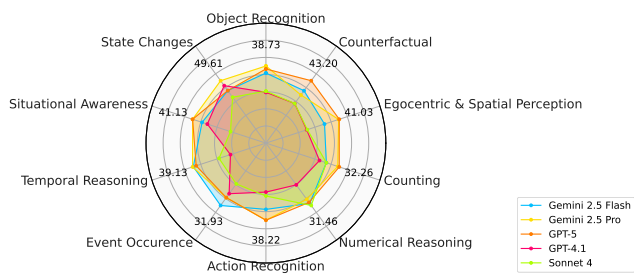


Figure 6. Accuracy of SotA models by skill