# VideoEspresso: A Large-Scale Chain-of-Thought Dataset for Fine-Grained Video Reasoning via Core Frame Selection

Supplementary Material

## A. Details of VideoEspresso

| Benchmark | Core Frames | CoT | # Questions |
|---|---|---|---|
| How2QA [21] | ✗ | ✗ | 2,852 |
| ActivityNet-QA [50] | ✗ | ✗ | 8,000 |
| NExT-QA [43] | ✗ | ✗ | 8,564 |
| MovieChat [36] | ✗ | ✗ | 13,000 |
| TVQA [15] | ✗ | ✗ | 15,253 |
| MSRVTT-QA [45] | ✗ | ✗ | 72,821 |
| VideoCoT [40] | ✗ | T | 11,182 |
| ***VideoEspreeso*** | ✓ | **T&V** | **203,546** |

Table 6. **Dataset comparison** between videoQA datasets. **T** and **V** represent the textual and visual elements in the CoT, respectively.

**Dataset Comparison.** Existing VideoQA datasets [15, 21, 36, 40, 43, 45, 50] are limited by manual annotations, making it challenging to scale up to meet the demands of LVLM training. In contrast, our proposed dataset, *VideoEspresso*, contains over 200K question-answer pairs (Tab. 6), significantly enhancing the dataset scale. Moreover, we annotate highly relevant core frames within the videos, providing a fine-grained representation of temporal information. While VideoCoT [40] only introduces text-level chains of thought (CoT), we address the gap in previous work by incorporating visual elements into CoT process.

**Experimental Setting.** To comprehensively evaluate the capabilities of LVLMs on VideoQA tasks, we selected: (1) closed-source large models, such as GPT-4o [31] and Qwen-VL-Max [3]; (2) general-purpose LVLMs that claim strong video capabilities on video benchmarks, such as InternVL [7] and Qwen2-VL [3]; and (3) popular video LVLMs, such as LongVA [52] and mPLUG-Owl3 [49].To ensure the fairness of the reported accuracies, the video frame sampling scheme, temperature, and other parameters follow the settings from the original paper. Additionally, we standardize the maximum token length of the outputs to 512. As our model training details, the learning rate is set to 2e-5, the warmup rate is 0.03, and we train the model for one epoch with global batch size of 16. The training and evaluation process is facilitated on 8 NVIDIA-A100 GPUs.

**Dataset details.** As shown in Tab. 7, *VideoEspresso* comprises 14 tasks, with the training and testing sets divided according to specific proportions. The detailed question design for each task is presented in Tab. 9. As shown in Fig. 11, traditional videoQA datasets sample all frames of a video at equal intervals. In contrast, *VideoEspresso* only focuses on the core frames of the video, which are highly relevant to the question. Unlike conventional videoQA tasks, which predominantly focus on querying actions or participants within the video, our dataset prioritizes the fine-grained logical reasoning, requiring a deeper understanding of complex temporal and contextual relationships. Moreover, the analysis of multimodal evidence integrated within the Chain-of-Thought reasoning process enhances both the accuracy and robustness of the generated answers, ensuring they are substantiated by comprehensive contextual understanding.

| Task | # Train Set | # Test Set |
|---|---|---|
| Causal Inference | 87,009 | 426 |
| Contextual Interpretation | 20,057 | 109 |
| Event Process | 29,227 | 174 |
| Interaction Dynamics | 7,322 | 62 |
| Behavior Profiling | 660 | 57 |
| Emotional Recognition | 3,505 | 65 |
| Influence Tracing | 5,749 | 72 |
| Role Identification | 9,134 | 63 |
| Narrative Structuring | 3,940 | 62 |
| Thematic Insight | 10,650 | 61 |
| Situational Awareness | 1,018 | 50 |
| Cooking Steps | 276 | 53 |
| Ingredient Details | 22,552 | 98 |
| Traffic Analysis | 1,065 | 30 |
| Total | 202,164 | 1,382 |

Table 7. **Tasks distribution** and **dataset split** in *VideoEspresso*.

**Human check on data construction.** We ensured data quality with extensive human supervision during the pipeline construction. Specifically: 1) *QA Pair Construction*: We manually reviewed dataset styles, set frame intervals for captioning, and adjusted similarity thresholds. Prompts for QA construction were repeatedly refined; 2) *CoT Annotation*: We iteratively optimized prompts and parameters for GroundingDINO based on generated results, balancing human oversight with progressive automation. These measures ensured data quality and minimized noise.

**Effectiveness of open-source models.** The framework is compatible with open-source models and will not be corrupted. The OSS models demonstrate powerful capabilities to handle complex tasks (*e.g.* data construction). We incorporated the latest open-source model, DeepSeek-R1. While

Qwen2-72B generates questions based on straightforward factual reasoning, GPT-4o and DeepSeek-R1 generate more complex questions involving advanced reasoning, such as causal inference and event processes.

| config | Stage1 | Stage2 |
|---|---|---|
| input resolution | 224 | 224 |
| max token length | 6144 | 6144 |
| LoRA | True | |
| weight ratio | 0.02 | |
| learning rate schedule | cosine decay | |
| learning rate | 2e-5 | 1e-5 |
| batch size | 16 | |
| warmup epochs | 0.03 | 0.03 |
| total epochs | 1 | 1 |

Table 8. **Training Hyperparameters** for different stages.

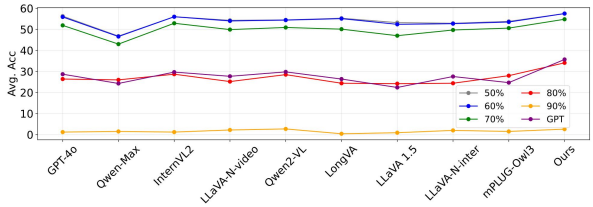


Figure 6. **Threshold-Performance Distribution.**

## B. Training Implementation

The hyperparameters used at different training stages are listed in Tab. 8, following LLaVA-Next architecture [17, 24]. During both stage, we leverage diverse instruction data and integrate LoRA modules [13] into the LLM with a rank of 16, an alpha value of 32, and a dropout rate of 0.1. Flash attention [8] is applied to accelerate the training process.

## C. Prompt Details

In this section, we present the complete set of prompts utilized in the data generation pipeline, alongside those employed for subjective evaluation. Specifically, these include the prompt designed for QA construction in Fig. 7, the prompt aimed at filtering low-quality QA pairs in Fig. 8, the prompt used for constructing CoT evidence in Fig. 9, and the prompt applied for subjective evaluation in Fig. 10.

## D. Evaluation Analysis

**Construction of Test set.** For all questions, we devised three distractor options that maintain consistent contextual relevance and similar linguistic structures to the correct answer while presenting distinct factual inaccuracies, thereby enhancing the robustness of the objective process. Furthermore, to mitigate potential biases arising from significant

**Prompt: QA Construction**

Please design diverse multi-image reasoning problem-answer pairs based on the description of the video frame sequence given below.
{captions}
Attention:
1. For each pair, please give the selected ORIGINAL captions in list form, questions, and answers in the order.
2. It is required that the designed questions MUST use multi-image information.
3. The designed questions must involve relatively complex reasoning. For example, by the cause of certain images, the result of certain images is obtained.
4. These images are all derived from the same video, so the similar objects between the different images are likely to be the same.
5. DO NOT generate questions that are highly subjective, including keywords such as: emotional, spiritual, contribution, importance, and implication.

Figure 7. **QA-Construction Prompt.**

**Prompt: QA Filter**

Use the provided criteria flag to assess each QA pair for quality.
Question: {question}
Answer: {answer}
For each QA pair flagged as low-quality, provide a brief explanation indicating which criterion was violated (e.g., "Subjective Question," "Lack of Continuity," "Overly Open-Ended Question").
Ensure high-quality QA pairs maintain alignment with the video's observable content, narrative flow, and context.

Figure 8. **QA-Filter Prompt.**

**Prompt: CoT evidence**

Please selected the most relative captions to the question from the caption list:
{captions}
{question}

And extract NO MORE THAN TWO key objects (noun form) for each caption in the captions according to the question and answer provided below, and combine these key elements into a complete sentence of evidence to explain the reasons for the answer.
{caption}
{answer}

Figure 9. **CoT-Evidence Construction Prompt.**

**Prompt: Subjective Evalutaion**

You are a scoring assistant for checking text quality. Please help me evaluate the following answers based on the question and the correct answer. \n
Question: {question}\n
Correct Answer: {reference_answer}\n
Model Output: {model_output}\n
Score each aspect from 1 to 10, including logic, factuality, accuracy, conciseness, and overall. For logic, evaluate how well the reasoning and structure of the response align with the question and whether the conclusions follow coherently (1-2: entirely illogical; 3-4: inconsistent or poorly structured; 5-6: partially logical with minor gaps; 7-8: mostly logical with rare issues; 9-10: fully logical and coherent). For factuality, assess the correctness of information and absence of errors (1-2: mostly incorrect or misleading; 3-4: significant factual inaccuracies; 5-6: some minor inaccuracies; 7-8: highly factual with rare errors; 9-10: entirely factual). For accuracy, consider the precision of the response in addressing the query (1-2: irrelevant or off-topic; 3-4: partially inaccurate; 5-6: moderately accurate; 7-8: accurate with minimal flaws; 9-10: perfectly accurate). For conciseness, rate the ability to convey the message effectively without unnecessary verbosity (1-2: excessively wordy or incomplete; 3-4: moderately verbose or lacks focus; 5-6: somewhat concise but could be improved; 7-8: mostly concise with rare verbosity; 9-10: perfectly concise and to the point). For overall, provide an integrated score reflecting the holistic quality of the response.
In the format of the dictionary (including brackets), return all your scoring results, ensuring your scores are integers, the format ONLY should be as follows:\n
For example: {'Logic': 9, 'Factuality': 6, 'Accuracy': 7, 'Conciseness': 7, 'Overall': 7}.\n

Figure 10. **Subjective Evaluation Prompt.**

| | |
|---|---|
| **Logical Reasoning** | |
| **Causal Inference** | *How did the actions of the robot and display on the screen contribute to the successful resolution in the control room?* |
| **Contextual Interpretation** | *How does the presence of the small cat and George's exploration relate to the chef's activities?* |
| **Event Process** | *What transition do the rabbits experience from the time the moon rose to when they drift off to sleep?* |
| **Social Understanding** | |
| **Interaction Dynamics** | *Considering the atmosphere and expressions depicted, what can be concluded about the progression of the interaction between the man and the woman?* |
| **Behavior Profiling** | *Discuss how the actions of the baby triceratops with different dinosaurs reveal aspects of its behavior and the responses of the other dinosaurs.* |
| **Emotional Recognition** | *How does the emotional journey of the small purple dinosaur from feeling lost to excitement tie into the group's decision to explore the cave?* |
| **Influence Tracing** | *How did the presence of the dolphin and the sea monster influence the dinosaurs' experience at the waterbody?* |
| **Discourse Comprehension** | |
| **Role Identification** | *How does the woman's role in coordinating town safety relate to the device's activation with a green checkmark and an orange flame?* |
| **Narrative Structuring** | *Considering the changes between the two frames, what can you infer about the narrative progression between the two depicted scenes?* |
| **Thematic Insight** | *How do the changing production logos contribute to the thematic preparation for the viewer before the main storyline begins?* |
| **Situational Awareness** | *Based on the sequence of events, how does the situation described contribute to the visual effect observed in the third frame?* |
| **Reality Application** | |
| **Cooking Steps** | *Considering the sequence of actions, what cooking technique is being employed, and how is it crucial for the fried chicken?* |
| **Ingredient Details** | *If the person is preparing chili con carne, what is the purpose of the liquid being poured into the pan?* |
| **Traffic Analysis** | *Analyze the potential destinations of the visible vehicles based on their types and cargo as inferred from the images.* |

Table 9. **Our proposed task categories with question prototypes.**


Figure 11. **Comparison** between *VideoEspresso* and other VideoQA dataset.

token-length disparities in the second step of the objective evaluation, we employed GPT-4o [31] to standardize the length and ensure a balanced distribution across all answer options for each question (shown in Fig. 13).

**Details of Objective Evaluation.** As illustrated in Algorithm 1, our objective evaluation is divided into two distinct steps. In the first step, the semantic similarity between the model's output $O$ and the reference answer $R$ is computed. If the similarity score $S_R$ falls below the predetermined threshold $tau = 80\%$, the output is deemed incor-

**Algorithm 1** Objective Evaluation for Open-Ended Output

**Require:** Model output $O$, reference answer $R$, threshold $\tau = 80\%$, distractors $\{D_1, D_2, D_3\}$
**Ensure:** Evaluation result: Correct or Incorrect
▷ *Step 1: Semantic Similarity Assessment*
Compute semantic similarity $S_R = \text{Sim}(O, R)$
**if** $S_R < \tau$ **then**
    **Return: Incorrect**
**end if**
▷ *Step 2: Confounding Distractor Analysis*
**for** each distractor $D_i$ in $\{D_1, D_2, D_3\}$ **do**
    Compute semantic similarity $S_{D_i} = \text{Sim}(O, D_i)$
    **if** $S_{D_i} > S_R$ **then**
        **Return: Incorrect**
    **end if**
**end for**
**Return: Correct**

rect. In the second step, a set of three carefully selected confounding distractors $\{D_1, D_2, D_3\}$ is introduced for each reference answer. The semantic similarity $S_{D_i}$ between the model's output and each distractor is then computed. If any distractor's similarity score $S_{D_i}$ exceeds $S_R$, the output is categorized as incorrect. Only outputs meeting the criteria in both steps are ultimately classified as correct.

**The Threshold of 80%.** The 80% threshold was determined experimentally. As shown in Fig. 6, extremely high thresholds inflate performance metrics, overestimating the model's capabilities, while low thresholds deflate metrics, leading to overly conservative assessments. Using GPT-4o for semantic evaluation, we found the 80% threshold aligns well with the evaluation curve, demonstrating its robustness while balancing accuracy and cost.

**Analysis of test set.** As depicted in Fig. 12, we present the example of reference answers and distractor options within the test set. The figure highlights factual inaccuracies in the distractor options using red annotations, while the correct answers are distinctly marked in green for clarity and emphasis. The token length disparities between reference answers and the longest distractor option, as shown in Fig. 13, predominantly are confined to the interval $[-10, +10]$, indicating that the disparity in length between correct answers lie within is relatively minor. The distribution shows near symmetry along the y-axis, indicating a balanced pattern: in about half of the cases, reference answers are longer than distractors, while in the remaining cases, distractors are longer.

## E. Case Study

Leveraging *VideoEspresso*, the LVLM demonstrates superior performance in fine-grained reasoning tasks for video understanding. As shown in Fig. 14, while GPT-4o generates seemingly rich and plausible answers, including detailed analyses of elephant and monkey behaviors, it incorporates a significant amount of video-irrelevant information, such as "*and their tusks for tasks like stripping bark or digging for roots*" and "*social foraging*". This diminishes the proportion of visually grounded outputs. This issue is more pronounced in Fig. 15, where GPT-4o performs extensive analysis based on erroneous reasoning due to its failure to account for fine-grained inter-frame relationships, e.g., "*If the jacket is visibly marked by soot, burns, or other damage, it suggests the individual was close to the fire.*" From these examples, we underscore the critical role of visual information in video reasoning QA tasks.

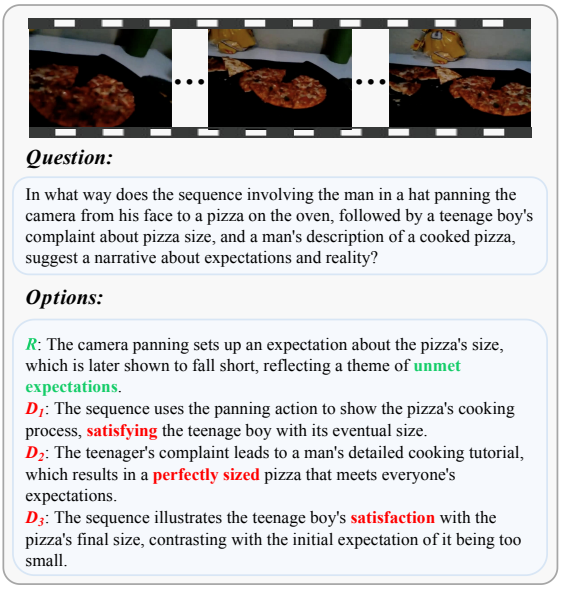


Figure 12. **Example of test set.** $R$ represent the Reference Answer, while $D_i$ stand for the $i$-th Distractor.

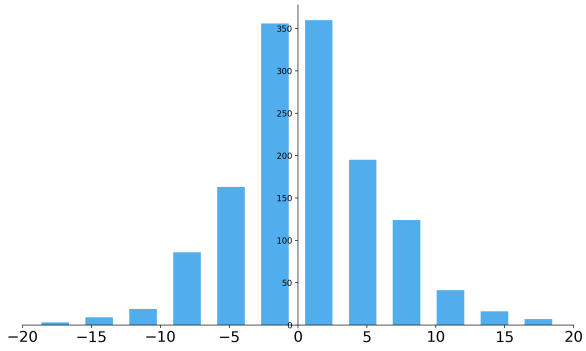


Figure 13. **The Distribution** of token length disparities between reference answers and the longest distractor option.

How do the physical characteristics of elephants and monkeys as portrayed in the video affect their foraging behaviors and strategies in their respective environments?

**GPT-4o:**
The physical adaptations of elephants and monkeys play a crucial role in shaping their foraging behaviors and strategies. **Elephants utilize their trunks for dexterous tasks like reaching high branches and gathering food**, their large size to access vegetation out of reach for others, **and their tusks for tasks like stripping bark or digging for roots ...** In contrast, **monkeys rely on their agility to climb trees and access canopy resources**, **opposable thumbs** to manipulate and peel fruit, and **social foraging** to locate food and ensure safety. These traits align with their environments, enabling efficient resource use and survival.

***VideoEspresso:***
The physical characteristics of elephants and monkeys play a significant role in their foraging behaviors and strategies. In the image of the elephant, **its large size and strong trunk enable it to reach high branches for food**, suggesting that it uses its strength and dexterity to access resources that are otherwise inaccessible to smaller animals. This adaptation allows it to exploit a niche in its environment efficiently. On the other hand, the monkey's **agility and ability to navigate through dense palm fronds suggest that it uses its quick reflexes and nimble movements** to avoid predators and efficiently search for food in its environment. The monkey's physical characteristics allow it to **move quickly and precisely**, which is crucial for survival in a dense, **tropical forest setting**.

Figure 14. **Example of over-analysis with GPT-4o.**

What inference can you make about the connection between the close-up of the individual's jacket and the two individuals standing in the room in the news report?

**GPT-4o:**
**... If the jacket is visibly marked by soot, burns, or other damage, it suggests the individual was close to the fire**, either as a victim, witness, or participant in extinguishing the flames. The close-up of the individual's jacket and the two individuals standing in the room appear to have no direct connection, as the images focus on separate aspects of the fire incident **without indicating a relationship** between them.

***VideoEspresso:***
The close-up of the individual's jacket, which is described as a light-colored jacket with a hood, suggests that this person **may be connected to the two individuals standing in the room**. This inference is based on the similarity in attire and the context provided by the news report about an accident involving a car. **The person wearing the light-colored jacket might be involved in the same incident or event**, potentially as a **witness or participant**, which explains why they are also featured in the news report alongside the other individuals.

Figure 15. **Example of Non-factual response with GPT-4o.**