VideoCoT: A Video Chain-of-Thought Dataset with Active Annotation Tool

Yan Wang[1], Yawen Zeng[2]*, Jingsheng Zheng[1], Xiaofen Xing[1], Jin Xu[1,3], Xiangmin Xu[1]
[1]South China University of Technology, Guangzhou, China
[2]ByteDance, Beijing, China
[3]Pazhou Lab, Guangzhou, China
ftwyan@mail.scut.edu.cn
{yawenzeng11, zhengjohnson0}@gmail.com
{xfxing, jinxu, xmxu}@scut.edu.cn

Abstract

Multimodal large language models (MLLMs) are flourishing, but mainly focus on images with less attention than videos, especially in sub-fields such as prompt engineering, video chain-of-thought (CoT), and instruction tuning on videos. Therefore, we try to explore the collection of CoT datasets in videos to lead to video OpenQA and improve the reasoning ability of MLLMs. Unfortunately, making such video CoT datasets is not an easy task. Given that human annotation is too cumbersome and expensive, while machine-generated is not reliable due to the hallucination issue, we develop an automatic annotation tool that combines machine and human experts, under the active learning paradigm. Active learning is an interactive strategy between the model and human experts, in this way, the workload of human labeling can be reduced and the quality of the dataset can be guaranteed. With the help of the automatic annotation tool, we strive to contribute three datasets, namely VideoCoT, TopicQA, TopicCoT. Furthermore, we propose a simple but effective benchmark based on the collected datasets, which exploits CoT to maximize the complex reasoning capabilities of MLLMs. Extensive experiments demonstrate the effectiveness our solution.

1 Introduction

With the emergence of ChatGPT[1], large language models (LLMs) have experienced unprecedented growth and have gradually expanded into the multimodal domain. Pioneers have explored multiple feasible paths around multimodal large models (MLLMs), such as training MLLMs from scratch (e.g. Kosmos-1 (Huang et al., 2023)), or bridging LLMs and vision modules (e.g. BLIP-2 (Li et al., 2023b)). Moreover, prompt engineering, chain-of-thought (CoT), and instruction tuning for multimodal LLMs are also flourishing. However, the majority of current research focuses on images, with video research (Deng et al., 2022; Zeng et al., 2022) remaining underdeveloped. For instance, Alayrac et al. (2022) employs a video understanding model to extract features, which are then inputted, while Ye et al. (2023) utilizes multiple frames of the video as input. Similarly, few researchers have devoted attention to sub-fields such as video prompt engineering (Li et al., 2023a; Zeng, 2022), and video instruction fine-tuning (Zhang et al., 2023b). We attribute this phenomenon to the fact that MLLMs are less mature than LLMs that solely rely on natural language input, and there are still numerous issues to be explored.

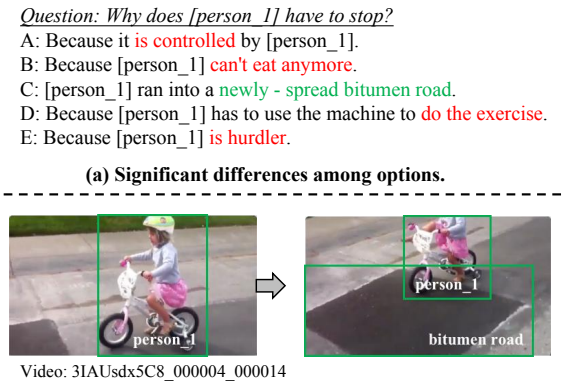


Figure 1: The case analysis of video question answering.

To advance the development of MLLMs for videos, our primary interest lies in CoT in videos. Video CoT has multiple benefits as follows: 1) Towards OpenQA in video. Currently, the VideoQA dataset widely adopts the form of multiple-choice questions, but there are significant differences between the answer options (Kamalloo et al., 2023). As illustrated in Fig.1(a), the options between A-E are significantly different, especially the descriptions of eating and being a hurdler are completely irrelevant to the video. This fact lead to models

*Corresponding author.
[1]https://openai.com/blog/chatgpt

Proceedings of the 3rd Workshop on Advances in Language and Vision Research, pages 92–101
August 16, 2024

Dataset	Rationale	Language	#Videos	#Q	Video Source	Annotation	QA Task
MSVD-QA (Chen and Dolan, 2011)	✗	English	1.9K	50K	Web Videos	Auto	OE
MovieQA (Tapaswi et al., 2016)	✗	English	6.7K	6.4K	Movies	Manual	MC
MSRVTT-QA (Xu et al., 2017)	✗	English	10K	243K	Web Videos	Auto	OE
TVQA (Lei et al., 2019)	✗	English	21K	152K	TV	Manual	MC
ActivityNet-QA (Yu et al., 2019)	✗	English	5.8K	58K	Web Videos	Manual	OE
NExT-QA (Xiao et al., 2021)	✗	English	5.4K	52K	YFCC-100M	Manual	MC,OE
Causal-VidQA (Li et al., 2022)	✗	English	26K	107K	Kinetics-700	Manual	MC
FIBER (Castro et al., 2022)	✗	English	28K	2K	VaTEX	Manual	OE
VideoCoT (Ours)	✓	English, Chinese	11K	22K	Kinetics-700	Auto, Manual	MC, OE
TopicQA (Ours)	✗	English, Chinese	11K	22K	Kinetics-700	Auto, Manual	MC, OE
TopicCoT (Ours)	✓	English, Chinese	11K	22K	Kinetics-700	Auto, Manual	MC, OE

Table 1: Comparision between our collected datasets (i.e. VideoCoT, TopicQA and TopicCoT) and other existing datasets. Among them, MC in the "QA Task" column means multiple-choice, while OE represents open-ended question answering.

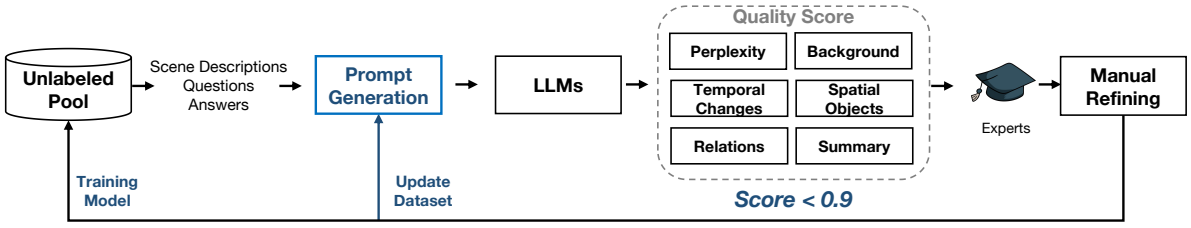
finding shortcuts to the dataset pattern. 2) Enhance understanding. Videos contain more temporal and spatial changes than images, and CoT can help capture the complex semantics of these changes (Zeng et al., 2021). As shown in Fig.1(b), the key to solving the question, that is, the girl changes from moving to stopping (temporal) and the appearance of the bitumen road (spatial), is to develop with the video. 3) Improving the reasoning ability of MLLMs. A more logical CoT can enhance the reasoning ability of MLLMs when used for training.

Although video CoT shows great potential, creating a video CoT dataset is a non-trivial task. The process of fully annotating CoTs by humans is both tedious and expensive, which is why we aim to develop an automatic pipeline for generating CoTs. Intuitively, one widely adopted strategy is to use off-the-shelf MLLMs or LLMs as assistants for reasoning. However, there are several challenges that need to be addressed. Firstly, MLLMs do not possess strong reasoning abilities and cannot directly generate reliable CoTs. Secondly, while LLMs have reasoning capabilities, they cannot use images as input for CoT generation. Lastly, machine-generated data is often unreliable due to ethical doubts and hallucination issues (Liu et al., 2023; Qin et al., 2023), which require human correction for quality control.

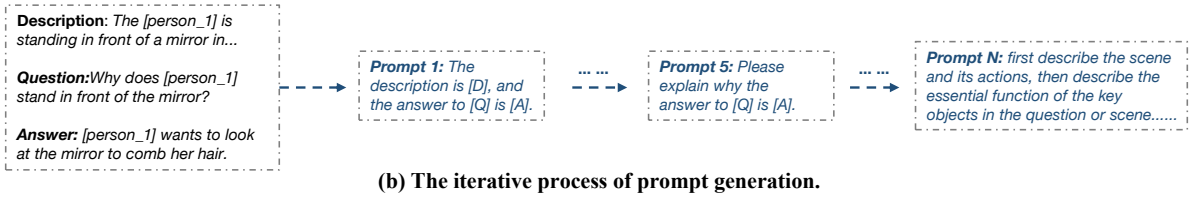
Therefore, in this paper, we develop an automatic annotation tool that combines machine and human experts, under the active learning paradigm (Zhang et al., 2023a). As shown in Fig.2, active learning is a strategy that involves interaction between the model and human experts, where the model actively seeks the opinions and standards of experts when encountering difficult samples (Zhai et al., 2022). In this way, the workload of human labeling can be reduced and the quality of the dataset can be guaranteed in the process(Wu et al., 2024; Lu et al., 2021). Specifically, we will train a prompt generator to guide LLMs to generate complex CoT based on video information. Meanwhile, we will formulate a quality score to evaluate the generated CoT sentences from multiple aspects. Among them, low-quality sentences will be modified by human experts, and the modified CoT will be used to train the prompt generator to guide LLMs to generate more reasonable CoT(Guo et al., 2022; Liu et al., 2022).

With the help of the aforementioned automatic annotation tools under the active learning paradigm, we strive to contribute three videoCoT datasets, namely **VideoCoT, TopicQA, TopicCoT**. Among them, VideoCoT is designed to supplement CoT between question and answer from existing datasets. Furthermore, we leverage the topic items in the dataset to construct TopicQA, which enables MLLMs to learn the relevant relationship between videos and topics, and TopicCoT, which facilitates reasoning about the topic relevance. Furthermore, we apply these datasets to propose a simple benchmark. Extensive experiments demonstrate the effectiveness of our datasets and solution. The main contributions are summarized as follows:

- To the best of our knowledge, this is the first work that introduces an automatic annotation tool under the active learning paradigm for complex CoT generation in the video domain.
- We have collected three dataset to fill the vacuum of Video CoT via our automatic annotation tool, namely VideoCoT, TopicQA, TopicCoT.
- We propose a simple but effective benchmark



(a) The process of active annotation tool.



(b) The iterative process of prompt generation.

Figure 2: The process of automatic dataset construction for VideoCoT and TopicCoT.

based on the collected datasets, which exploits CoT to achieve better reasoning ability.

2 Related Work

2.1 Multimodal Large Models

As a result of the flourishing development of LLMs (Pan and Zeng, 2023), many frameworks and techniques have been extended, such as prompt engineering, chain-of-thought, and instruction tuning. In the field of multimedia, these hotspots are still the topic of discussion (Li et al., 2024). Subsequently, Zhu et al. (2023) proposed mini-GPT4, Li et al. (2023b) introduced blip2, and Ye et al. (2023) intruduced mPLUG-OWL. However, the majority of current research focuses on images, with video research remaining underdeveloped. To fill the academic vacuum, we propose an automatic annotation tool under the active learning paradigm, and further collect three datasets based on it. In this way, the complex reasoning ability of MLLMs is improved (Rajesh et al., 2023; Zeng et al., 2024).

2.2 Chain-of-Thought

Chain-of-Thought (CoT) has been proven to be an effective strategy to enhance reasoning, and its effectiveness has been widely demonstrated in the field of LLMs (Ma et al., 2023). In the field of multimedia, works such as ScienceQA (Lu et al., 2022) and VisualCoT (Rose et al., 2023) have also been proposed. Inspired by the above work, we try to extend the potential of CoT in the field of video understanding, which helps improve the reasoning ability of MLLMs.

3 Dataset Collection

Following Causal-VidQA (Li et al., 2022), we built three datasets around videos based on Kinetics-700, namely VideoCoT, TopicQA, and TopicCoT. In this section, we will introduce the process of active annotation tool, on which both VideoCoT and TopicCoT are collected.

3.1 Active Annotation Tool

Fig.2 illustrates the pipeline of our automatic dataset construction approach, which implements the prompt generation for LLMs under the active learning paradigm to generate the logical CoT processes. Active learning is an interrogation method between the model and human experts (Zhang et al., 2023a), which reduces the annotation workload and guarantees the quality of the dataset.

Specifically, the automated process is divided into three steps, namely prompt generation, automatic scoring, and expert refinement. Among them, prompt generation aims to generate suitable prompt to guide LLMs to generate comprehensive and reasonable CoT, while automatic scoring checks the quality of machine-generated CoT from multiple quality dimensions. Among them, the low-quality CoT will be refined and modified by experts, which is also used to train the prompt generator to improve the quality of CoT generation.

3.1.1 Prompt Generation

We try to drive the off-the-shelf LLMs (i.e. GPT-4) to generate some high-quality CoT data for us, but unfortunately, the logic of the generated sentences obtained by the fixed template (i.e. prompt) is incomplete and incoherent. Therefore, we introduce a prompt generator to maximize the potential of

guiding LLMs and ultimately reduce manual labor.

Specifically, we borrow a summarization model (Rao et al., 2021) capable of handling long sentences as the prompt generator, which will be trained in interaction with human experts. In the initial stage, it is fed a long video description, a question and a answer, and finally outputs a short summary. Obviously, such a prompt is difficult to guide LLMs to get a reasonable CoT between the question and the answer, so it needs to learn from human modified sentences. We will present the scoring mechanism and human refinement in the next subsection.

After multiple rounds of iterations, the generator will flexibly deal with different videos to generate corresponding prompts. Thereafter, since MLLMs do not yet have good reasoning capabilities (which is what we hope to do), we still implement generation based on LLMs (i.e. GPT-4). Finally, after manual inspection with less labor, a reasonable CoT can be obtained, as shown in the Fig.3.

3.1.2 Automatic Scoring

In order for a quality-required CoT to be generated, we believe that a high-quality CoT $\mathcal{C}_{vCoT}$[2] should have both: 1) the generated sentences are fluent, 2) a comprehensive understanding of objects and relations, 3) and reasonable reasoning between the question and the answer. To achieve this, we design a scoring function $\mathcal{S}_{vCoT}$ that automatically evaluates from six dimensions, i.e., perplexity $\mathcal{S}_{ppl}$, background $\mathcal{S}_{bac}$, temporal changes $\mathcal{S}_{tem}$, spatial objects $\mathcal{S}_{spa}$, relations $\mathcal{S}_{rel}$, summary $\mathcal{S}_{sum}$.

$$\mathcal{S}_{vCoT} = \mathcal{S}_{ppl} + \mathcal{S}_{bac} + \mathcal{S}_{tem} + \mathcal{S}_{spa} + \mathcal{S}_{rel} + \mathcal{S}_{sum}. \tag{1}$$

Among them, the "perplexity" evaluates the fluency of generated CoT, and its reciprocal is used as part of the quality score (Basu et al., 2021). This score is closer to 1 when the CoT sentence $\mathcal{C}_{vCoT}$ is more fluent.

$$\mathcal{S}_{ppl} = \frac{1}{PPL(\mathcal{C}_{vCoT})}. \tag{2}$$

The "background" $\mathcal{S}_{bac}$ indicates whether the generated CoT describes the video scene or not. We collect some keywords to evaluate this, i.e., when a sentence of CoT has words such as *background*, *video scene*, etc., it is considered to meet the quality requirement.

$$\mathcal{S}_{bac} = \begin{cases} 1 & \text{if the video scene is described in } \mathcal{C}_{vCoT} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

The "spatial objects" $\mathcal{S}_{spa}$ and "temporal changes" $\mathcal{S}_{tem}$ represent how many objects and actions are included in the generated CoT, respectively. The objects and actions (extracted by GRiT(Wu et al., 2022)) that should be included are taken as the evaluation criteria, i.e. the more objects and actions are included in $\mathcal{C}_{vCoT}$, the higher the score $\mathcal{S}_{spa}$ and $\mathcal{S}_{tem}$. Conversely, if irrelevant objects or actions appear in the sentence $\mathcal{C}_{vCoT}$ (most likely hallucinations), the score will be negative.

$$\mathcal{S}_{spa} = \frac{\text{pos}_o(\mathcal{C}_{vCoT}) - \text{neg}_o(\mathcal{C}_{vCoT})}{\text{ground\_truth}(\mathcal{C}_{vCoT})}, \tag{4}$$

$$\mathcal{S}_{tem} = \frac{\text{pos}_a(\mathcal{C}_{vCoT}) - \text{neg}_a(\mathcal{C}_{vCoT})}{\text{ground\_truth}(\mathcal{C}_{vCoT})}, \tag{5}$$

where pos_o and pos_a indicate the number of objects and actions present in the CoT, where pos indicates real presence in the video, and neg indicates hallucinated objects or actions.

The "relations" $\mathcal{S}_{rel}$ represents whether the generated CoT has the analysis of spatio-temporal relationship among objects, and the connection with video scene. And the "summary" $\mathcal{S}_{sum}$ evaluates whether a summary is included in the generated $\mathcal{C}_{vCoT}$ (i.e., the answer is output via step-by-step

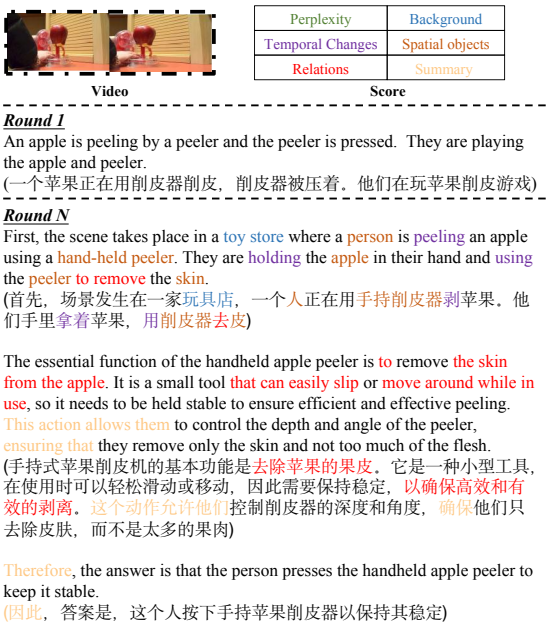


Figure 3: After multiple rounds of training, the quality score of the generated CoT is improved from 0.07 to 0.97.

[2] C_v represents "video", while C_{vCoT} represents Video-CoT, which serves to differentiate it from TopicCoT C_{tCoT}.

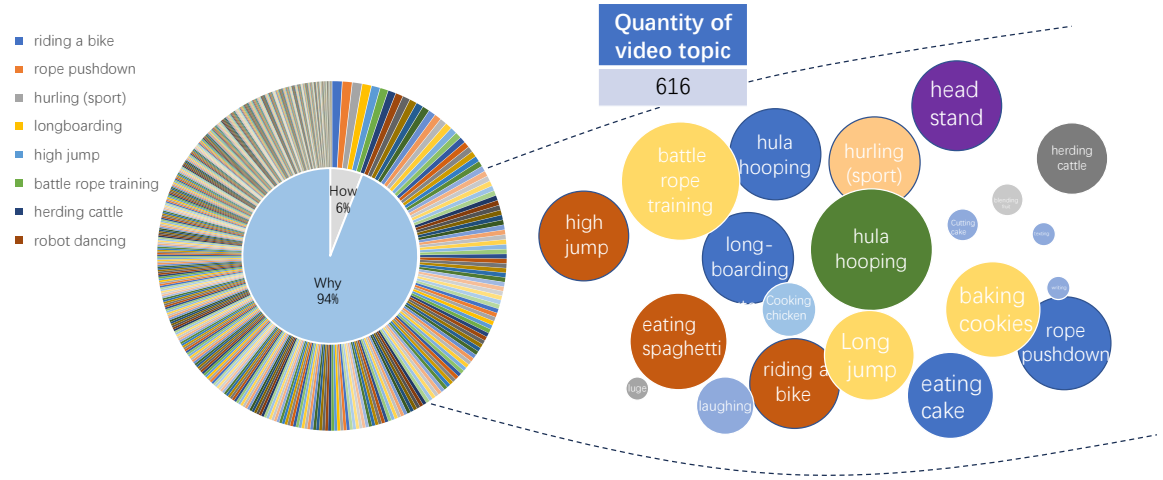


Figure 4: The topic and question distribution for VideoCoT and TopicCoT.

reasoning).

$$\mathcal{S}_{rel} = \begin{cases} 1 & \text{if the analysis is included in } \mathcal{C}_{vCoT} \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

$$\mathcal{S}_{sum} = \begin{cases} 1 & \text{if the summary is included in } \mathcal{C}_{vCoT} \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

All the above scores belong to the interval from 0 to 1, which is convenient for us to do further normalization. The automatic score S serves as a "rough indicator" to identify the worst sample and help us optimize prompt generator. In particular, since $\mathcal{S}_{spa}$ and $\mathcal{S}_{tem}$ are more important for this task, we set the balance parameters in Eqn.1 as $(0.1, 0.1, 0.3, 0.3, 0.1, 0.1)$. Furthermore, to control the quality of CoT, when the normalized score is lower than 0.9, it will be sent to human experts for refinement.

3.1.3 Expert Refinement

We enlisted ten human experts with backgrounds in artificial intelligence to participate in the annotation process. To ensure consistency in the labeling results across different experts, a 5-rounds pre-annotation training was conducted prior to official annotation. Specifically, each expert was required to label a small number of samples to gain an understanding of the annotation rules, which were standardized to ensure consistency among all participants.

For the generated CoT whose quality score is less than the threshold (i.e. 0.9), they will be modified by human experts. As much as possible, experts are asked to make sentences include scene descriptions in video, spatio-temporal relationships, and logical reasoning between the question and answer. Meanwhile, the refined samples will return to the dataset pool and participate in training of prompt generation until the quality of all annotations meets our requirements. Through this interactive active learning paradigm, the high-quality CoT are semi-automatically constructed.

3.2 Automatic Datatset Construction

With the help of the aforementioned annotation tool under the active learning paradigm, we strive to contribute three datasets, namely VideoCoT, TopicQA, TopicCoT.

3.2.1 VideoCoT

VideoCoT is designed to supplement CoT between question and answer from existing datasets, Causal-VidQA. Based on the settings, we collect $11,182$ samples containing CoT, as shown in Table 1.

3.2.2 TopicQA

Further, we leverage the topic items in the Kinetics-700 dataset to construct TopicQA, which enables MLLMs to learn the relevant relationship between videos and topics. In this dataset, we take "is the video relevant to the topic" as the question and "yes" or "no" as the answer.

3.2.3 TopicCoT

TopicCoT, similar to the construction process of VideoCoT, which contains step-by-step reasoning between questions and answers in TpoicQA. Specifically, TopicCoT $\mathcal{C}_{tCoT}$ is still based on our automatic annotation tool, but the scoring function is different, which is defined as follows:

$$\mathcal{S}_{tCoT} = \mathcal{S}_{ppl} + \mathcal{S}_{tem} + \mathcal{S}_{spa} + \mathcal{S}_{con} + \mathcal{S}_{sum}. \tag{8}$$

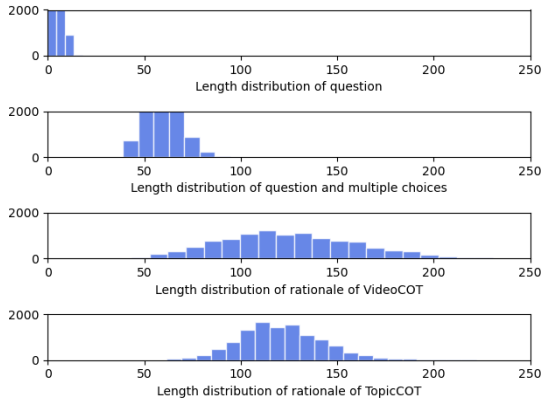


Figure 5: The length distribution of our dataset, where the y-axis represents the number of samples whose length is the x-axis value.

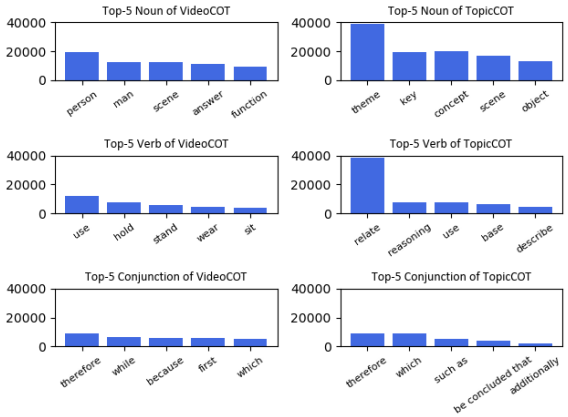


Figure 6: The top words of our dataset, where the y-axis represents the frequency of word count.

where $\mathcal{S}_{con}$ represents the concept of the topic, and the others are consistent with Eqn.1. Moreover, the balance parameters are set to $(0.1, 0.2, 0.2, 0.4, 0.1)$ for normalization. Then, when this score $\mathcal{S}_{tCoT}$ is less than 0.9, it will be sent to humans for modification.

3.3 Dataset Analysis

3.3.1 Property Quality

The statistical analysis of textual description in our VideoCoT and TopicCoT dataset is shown in Fig.5. Based on statistical results, the original dataset, which includes both questions and multiple choices, has an average length of approximately 50 words. In contrast, the rationale length of our VideoCoT and TopicCoT is distributed between 100 and 150 words.

3.3.2 Diversity Quality

To assess the diversity of sentences in the VideoCoT and TopicCoT datasets, we conduct a word frequency analysis of nouns, verbs, and conjunctions, which represent descriptive, temporal, and logical aspects, respectively. Fig.6 illustrates the top 5 frequency of each category in the rationale of the two datasets. **1) Noun**: We observe that the high-frequency nouns in VideoCoT mostly refer to specific objects, such as “person" and “man", as well as key words in the reasoning process, such as “scene", “answer" and “function". In contrast, the top nouns in TopicCoT mainly involve “topic" and “concept", indicating that detailed descriptions revolve around the topic and object concepts of the video. **2) Verb**: The main verbs in VideoCoT describe specific human activities, focusing on the temporal aspect of the video. In TopicCoT, the high-frequency verbs are mostly reasoning verbs, focusing on the association between the question and the topic of the video. **3) Conjunction**: The conjunction with the highest frequency in both datasets is “therefore", which indicates the logical and summary aspects of the rationale.

3.3.3 Visualization Quality

To verify the rationality of the human experts' operation, we also check some cases as shown in the Fig.3. There are two languages present in our dataset, namely English and Chinese. The initial generated by LLMs was of low quality, which hindered the establishment of relationships. However, after undergoing multiple round of interaction between human and model, the score of generated CoT increased from 0.07 to 0.97 points, indicating a significant improvement in the quality of the output.

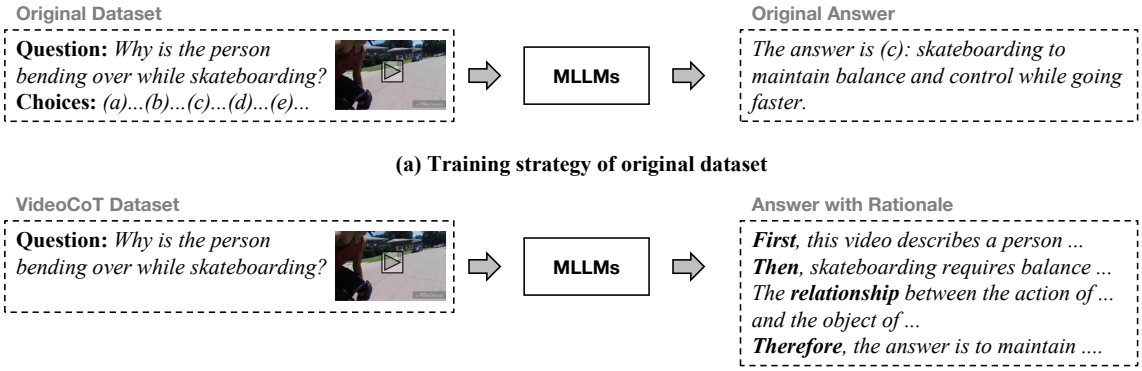
4 Proposed Method

The overall training framework is depicted with an illustration in Fig.7. For the task of video question answering (Zhong et al., 2022), multiple choice (MC) is more popular, but the differences between the options are too significant, and it is easy for the model to find shortcuts. Therefore, we are committed to achieving a free-form open-ended (OE) with logic rationale (Lu et al., 2022).

4.1 Training strategy of original dataset

The input of MC strategy is defined as $X = (X_Q, X_{MC}, X_V)$, where X_Q represents the question, X_{MC} represents answer options, and X_V represents the image.

Following the work of (Kamalloo et al., 2023), who trains the model using fixed long sentence



(a) Training strategy of original dataset

(b) Training strategy of our VideoCoT

Figure 7: Comparison of training strategies on the original dataset and our datasets.

templates with correct options for filling in the blanks, the probability of generating an answer can be formulated as follows:

$$p(Y|X_Q, X_{MC}, X_V) \tag{9}$$

$$=\sum_{t=1}^{m} \log p(y_t|y_{<t}, X_Q, X_{MC}, X_V), \tag{10}$$

where $Y=(y_1, y_2, \ldots, y_m)$ represents the target tokens.

4.2 Training strategy of VideoCoT

Similarly, the input of OE strategy is defined as $X=(X_Q, X_V)$.

In this way, the input X will be removed the answer options X_{MC}, while the target answer Y will be redefined as the rationale $R=(r_1, r_2, \ldots, r_n)$.

Formally, the probability of generating rationale can be formulated as follows:

$$p(R|X_Q, X_V)=\sum_{t=1}^{n} \log p(r_t|r_{<t}, X_Q, X_V). \tag{11}$$

Through this training strategy of CoT, more prior knowledge of MLLMs can be invoked, and finally answer questions through logical reasoning.

5 Experiments

5.1 Experimental Settings

5.1.1 Datasets

Our datasets are split into 3 non-overlapping subsets, where 0.6, 0.2 and 0.2 are used for training, validation and testing.

5.1.2 Evaluation Protocol

We adopt accuracy as our evaluation metric, which is utilized to measure whether the answers generated by models are correct. Notably, in the multi-choice setting, the accuracy Acc_{MC} can be directly compared with ground-truth. In the case of open-ended QA, we adopt two metrics, 1) Acc_{OE}(keywords): whether the "summary" sentence hits the keywords in the ground-truth answer. Specifically, keywords and their synonyms are acquired by giving some few-shot template and QA pair to GPT4. We then calculate the correct proportion of keywords for each question as its score. 2) Acc_{OE}(GPT-4): regard GPT-4[3] as a referee to evaluate semantic relevance.

5.1.3 Baselines

We select the following models as our baselines: mPLUG-Owl (Ye et al., 2023), VisualGLM (Du et al., 2022), mini-GPT4 (Zhu et al., 2023).

5.2 Overall Performance Comparison

To verify the effectiveness of our datasets, we train several MLLMs with the original dataset and our datasets respectively[4]. Among them, for the evaluation of OE task, we adopt two kinds of metrics, namely a hard metric (based on keywords) and a soft metric (based on GPT-4).

The experimental results are presented in Table 2, and the following observations can be made: 1) In comparison to the multi-choice setup, both models exhibit improved performance in open-ended QA accuracy. Upon analyzing the multi-choice outputs, it is evident that the models often provide justifications for each individual option rather than selecting a single response to address the given question. 2) The superiority of both VideoCoT trained MLLMs over the original method is evident in the improvements observed across both keyword

[3] https://openai.com/product/gpt-4

[4] TopicQA is an ordinary QA dataset, which will not be adopted to discuss the impact of CoT on reasoning ability, but it can still be a traditional QA dataset.

Model	Acc_{MC}	VideoCoT		TopicCoT	VideoCoT & TopicCoT
		Acc_{OE}(GPT-4)	Acc_{OE}(keywords)	Acc_{OE}(GPT-4)	Acc_{OE}(GPT-4)
mPLUG-Owl	31.51%	48.32%	52.66%	40.12%	–
VisualGLM	13.81%	45.32%	46.78%	23.34%	–
mini-GPT4	29.05%	43.58%	51.21%	19.21%	–
mPLUG-Owl (trained)	–	77.42% (+29.1)	81.24% (+28.58)	89.76% (+49.64)	90.18%
VisualGLM (trained)	–	69.91% (+24.59)	70.71% (+23.93)	78.96% (+55.62)	79.24%
mini-GPT4 (trained)	–	64.14% (+20.56)	75.20% (+23.99)	82.55% (+63.34)	82.85%

Table 2: Overall performance comparison among various methods on our VideoCoT and TopicCoT.

and GPT-4 metrics. This highlights the significant impact of employing a chain of thoughts within the generation model's creative process. 3) We also observe that the accuracy of keywords on all models surpasses the accuracy of GPT-4, which is due to the former metric being more relaxed than the latter. 4) Additionally, we conduct an experiment utilizing a hybrid training dataset comprising both VideoCoT and TopicCoT. The subsequent evaluation of models take place on the testing of VideoCoT. Remarkably, when contrasted with models solely trained on VideoCoT, the GPT-4 metric exhibited a noteworthy improvement through hybrid training. This improvement surpassed the performance of all models that are only trained on VideoCoT. This outcome serves as a compelling indicator that hybrid training fosters a reciprocal influence, allowing models to acquire the capacity for incremental and reasoned thinking.

5.3 Reasoning Ability Visualization

The visualization is shown in Fig.8, the mPLUG-Owl possesses the capability to depict the content of the image and execute the basic task of question and answer. However, its performance is unsatisfactory when confronted with more complex questions that necessitate reasoning. Conversely, upon being trained on our datasets, it acquires the ability to identify objects in the image (e.g. "a group of people"), discern the fundamental functions of objects or events (e.g. "the essential fuction of"), and finally integrate objects and relationships to engage in reasoning (e.g. "because they might participating in a fitness event").

6 Conclusions

In this work, we strive to explore the collection of CoT datasets on videos to bootstrap OpenQA on videos and improve the inference ability of MLLMs. To reduce the cost of manual annotation, we develop an automatic annotation tool that combines machine and human experts, under the active learning paradigm. With the help of this annotation tool, we contribute three videoCoT datasets, namely VideoCoT, TopicQA, TopicCoT. Experimental results show that our datasets achieve superior effectiveness, diversity and explainability.

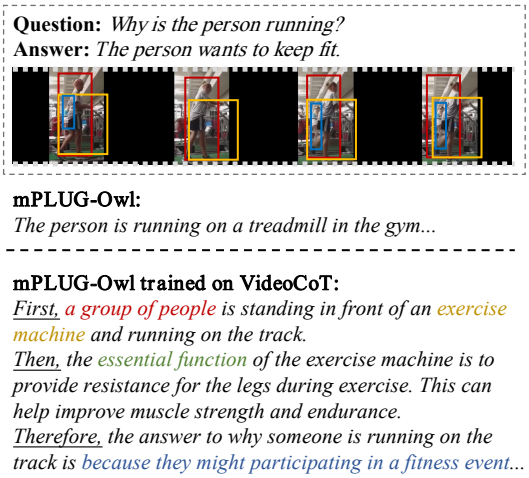


Figure 8: The visualization case of generated answers.

Acknowledgements

This work is supported in part by the National Natural Science Foundation of China (62372187), in part by the National Key Research and Development Program of China (2022YFC3601005) and in part by the Guangdong Provincial Key Laboratory of Human Digital Twin (2022B1212010004).

Limitations

In regards to the active annotation tool, using our tool on additional datasets can enhance the visual reasoning abilities of more models. However, funding constraints limited the invitation of annotation experts. Nonetheless, we are committed to expanding the impact of this paper in future research. Moreover, our training resources currently restrict the application of our dataset to significantly more larger models.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: a visual language model for few-shot learning. *Preprint*, arXiv:2204.14198.

Sourya Basu, Govardana Sachitanandam Ramachandran, Nitish Shirish Keskar, and Lav R. Varshney. 2021. Mirostat: A neural text decoding algorithm that directly controls perplexity. *Preprint*, arXiv:2007.14966.

Santiago Castro, Ruoyao Wang, Pingxuan Huang, Ian Stewart, Oana Ignat, Nan Liu, Jonathan C. Stroud, and Rada Mihalcea. 2022. Fiber: Fill-in-the-blanks as a challenging video understanding evaluation framework. *Preprint*, arXiv:2104.04182.

David L. Chen and William B. Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *ACL*, Portland, OR.

Chaorui Deng, Qi Wu, Qingyao Wu, Fuyuan Hu, Fan Lyu, and Mingkui Tan. 2022. Visual grounding via accumulated attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3):1670–1684.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Yong Guo, Yin Zheng, Mingkui Tan, Qi Chen, Zhipeng Li, Jian Chen, Peilin Zhao, and Junzhou Huang. 2022. Towards accurate and compact architectures via neural architecture transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6501–6516.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, Qiang Liu, Kriti Aggarwal, Zewen Chi, Johan Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, and Furu Wei. 2023. Language is not all you need: Aligning perception with language models. *Preprint*, arXiv:2302.14045.

Ehsan Kamalloo, Nouha Dziri, Charles L. A. Clarke, and Davood Rafiei. 2023. Evaluating open-domain question answering in the era of large language models. *Preprint*, arXiv:2305.06984.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L. Berg. 2019. Tvqa: Localized, compositional video question answering. *Preprint*, arXiv:1809.01696.

Jiangtong Li, Li Niu, and Liqing Zhang. 2022. From representation to reasoning: Towards both evidence and commonsense reasoning for video question-answering. In *CVPR*.

Juncheng Li, Kaihang Pan, Zhiqi Ge, Minghe Gao, Hanwang Zhang, Wei Ji, Wenqiao Zhang, Tat-Seng Chua, Siliang Tang, and Yueting Zhuang. 2023a. Empowering vision-language models to follow interleaved vision-language instructions. *Preprint*, arXiv:2308.04152.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *Preprint*, arXiv:2301.12597.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2024. Videochat: Chat-centric video understanding. *Preprint*, arXiv:2305.06355.

Jing Liu, Bohan Zhuang, Zhuangwei Zhuang, Yong Guo, Junzhou Huang, Jinhui Zhu, and Mingkui Tan. 2022. Discrimination-aware network pruning for deep model compression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4035–4051.

Xiao Liu, Hanyu Lai, Hao Yu, Yifan Xu, Aohan Zeng, Zhengxiao Du, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. Webglm: Towards an efficient web-enhanced question answering system with human preferences. *Preprint*, arXiv:2306.07906.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Preprint*, arXiv:2209.09513.

Yaojie Lu, Hongyu Lin, Jin Xu, Xianpei Han, Jialong Tang, Annan Li, Le Sun, Meng Liao, and Shaoyi Chen. 2021. Text2event: Controllable sequence-to-structure generation for end-to-end event extraction. *Preprint*, arXiv:2106.09232.

Yuhan Ma, Haiqi Jiang, and Chenyou Fan. 2023. Scicot: Leveraging large language models for enhanced knowledge distillation in small models for scientific qa. *Preprint*, arXiv:2308.04679.

Keyu Pan and Yawen Zeng. 2023. Do llms possess a personality? making the mbti test an amazing evaluation for large language models. *Preprint*, arXiv:2307.16180.

Yujia Qin, Zihan Cai, Dian Jin, Lan Yan, Shihao Liang, Kunlun Zhu, Yankai Lin, Xu Han, Ning Ding, Huadong Wang, Ruobing Xie, Fanchao Qi, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2023. Webcpm: Interactive web search for chinese long-form question answering. *Preprint*, arXiv:2305.06849.

Kousik Rajesh, Mrigank Raman, Mohammed Asad Karim, and Pranit Chawla. 2023. Bridging the gap: Exploring the capabilities of bridge-architectures for complex visual reasoning tasks. *Preprint*, arXiv:2307.16395.

Roshan Rao, Joshua Meier, Tom Sercu, Sergey Ovchinnikov, and Alexander Rives. 2021. Transformer protein language models are unsupervised structure learners. In *ICLR*. OpenReview.net.

Daniel Rose, Vaishnavi Himakunthala, Andy Ouyang, Ryan He, Alex Mei, Yujie Lu, Michael Saxon, Chinmay Sonar, Diba Mirza, and William Yang Wang. 2023. Visual chain of thought: Bridging logical gaps with multimodal infillings. *Preprint*, arXiv:2305.02317.

Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2016. Movieqa: Understanding stories in movies through question-answering. *Preprint*, arXiv:1512.02902.

Danyang Wu, Zhenkun Yang, Jitao Lu, Jin Xu, Xiangmin Xu, and Feiping Nie. 2024. Ebmgc-gnf: Efficient balanced multi-view graph clustering via good neighbor fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–15.

Jialian Wu, Jianfeng Wang, Zhengyuan Yang, Zhe Gan, Zicheng Liu, Junsong Yuan, and Lijuan Wang. 2022. Grit: A generative region-to-text transformer for object understanding. *Preprint*, arXiv:2212.00280.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. Next-qa:next phase of question-answering to explaining temporal actions. *Preprint*, arXiv:2105.08276.

Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. 2017. Video question answering via gradually refined attention over appearance and motion. In *ACM MM*, MM '17, page 1645–1653, New York, NY, USA. Association for Computing Machinery.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. 2023. mplug-owl: Modularization empowers large language models with multimodality. *Preprint*, arXiv:2304.14178.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. 2019. Activitynet-qa: A dataset for understanding complex web videos via question answering. *Preprint*, arXiv:1906.02467.

Runhao Zeng, Wenbing Huang, Mingkui Tan, Yu Rong, Peilin Zhao, Junzhou Huang, and Chuang Gan. 2022. Graph convolutional module for temporal action localization in videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6209–6223.

Yawen Zeng. 2022. Point prompt tuning for temporally language grounding. In *SIGIR*, pages 2003–2007.

Yawen Zeng, Da Cao, Xiaochi Wei, Meng Liu, Zhou Zhao, and Zheng Qin. 2021. Multi-modal relational graph for cross-modal video moment retrieval. In *CVPR*, pages 2215–2224. Computer Vision Foundation / IEEE.

Yawen Zeng, Yiru Wang, Dongliang Liao, Gongfu Li, Jin Xu, Hong Man, Bo Liu, and Xiangmin Xu. 2024. Contrastive topic-enhanced network for video captioning. *Expert Systems with Applications*, 237:121601.

Yajing Zhai, Yawen Zeng, Da Cao, and Shaofei Lu. 2022. Trireid: Towards multi-modal person re-identification via descriptive fusion model. In *ICMR*, pages 63–71. ACM.

Zhisong Zhang, Emma Strubell, and Eduard Hovy. 2023a. A survey of active learning for natural language processing. *Preprint*, arXiv:2210.10109.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023b. Multimodal chain-of-thought reasoning in language models. *Preprint*, arXiv:2302.00923.

Yaoyao Zhong, Junbin Xiao, Wei Ji, Yicong Li, Weihong Deng, and Tat-Seng Chua. 2022. Video question answering: Datasets, algorithms and challenges. *Preprint*, arXiv:2203.01225.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *Preprint*, arXiv:2304.10592.