Commonsense Video Question Answering through Video-Grounded Entailment Tree Reasoning

Supplementary Material

7. Additional quantitative results

Comparison with state-of-the-art. In addition to the state-of-the-art comparison of VQA methods on the de-biased set (Tab. 3 in the main manuscript), we also provide the comparison results on the original test set. Our framework remains competitive with state-of-the-art VQA methods, though our reasoner is about $\mathbf{250}\times$ **smaller** in parameters than other methods. Moreover, it also achieves new state-of-the-art results in some cases, especially for temporal reasoning. We also notice that the superiority of our framework in the de-biased sets is more significant than in the original sets. This observation highlights the effectiveness of our framework in reasoning over joint visual-text information when the reliance on textual biases is mitigated.

Results on Env-QA. To further validate the effectiveness and generalizability of our framework, we also test on the Env-QA dataset [1], which mainly consists of ego-centric videos collected from virtual environments. We report the results under three types of questions (state, event, and order reasoning), focusing on temporal reasoning. Results are summarized in Tab. 11. We observe that incorporating our framework brings consistent improvement across the video- and image-based VLMs.

8. Quality assessment of de-biased set

We conducted a human evaluation to assess the quality of our de-biased set. Specifically, we randomly selected 1000 original QA samples and 1000 de-biased QA from the NeXT-QA dataset and presented them to four volunteers. The volunteers were required to select the best answer from all available options under two distinct conditions: (1) without watching the associated video content, and (2) with the video content available for reference. Results are summarized in Tab. 9. It can be seen that humans can reliably answer rewritten questions (94%), comparably to the original set (96%). Meanwhile, in the original set, humans confirm the textual biases and can achieve an accuracy of 79% without analyzing the video; yet, they cannot easily deduce the correct answer solely from the de-biased question-answer pairs (accuracy of 44%). Hence, our de-biased QA ensures all options pose a comparable level of commonsensical association rather than having a dominant association to the correct answer. It demonstrates that our de-biasing procedure retains the fairness of the benchmark while effectively reducing the textual shortcuts.

Method	Original set	De-biased set
Human w/o video	79.3	44.6
Human w/ video	96.4	93.9

Table 9. Results of subjective human evaluation for NeXT-QA, which are derived from the average accuracy of four volunteers.

9. Additional ablation studies

Design of anchor frame localization. In our implementation, we directly prompt an LLM to retrieve the anchor frame based on the structured representations of both the fact statement and candidate frames. Additionally, we test other available metrics for anchor frame localization, including (1) *visual-text similarity*: calculating frame-question similarity using CLIP; (2) *text-text similarity*: measuring the similarity between text embeddings of frame-wise captions and the question text; and (3) *LLM-evaluated relevance score*: following the Video-Tree approach [1], we prompt the LLM to assign a relevance score to each frame based on its caption and the question text. The comparison results, summarized in Tab. 12, show that our solution performs better than all competitors. Notably, the LLM-evaluated relevance score demonstrates comparable performance to our method, while traditional visual-text and text-text similarity metrics lag behind. This indicates that modern LLMs are highly effective and generalizable tools for approximate retrieval.

Modality for proving entailment. There is a growing trend of transforming multimodal tasks into text-only tasks by converting other modalities into text, enabled by generative multimodal models. This paradigm enables powerful LLMs to tackle challenging tasks more effectively. In our method, we also explore the reasoning paradigm of the prover, comparing our implementation with a purely text-based reasoning solution. Specifically, given captions of the visual evidence for each statement, we directly use an off-the-shelf LLM to assess the confidence score for each statement. The comparison results in Tab. 13 show that the text-only reasoning paradigm achieves comparable performance when a strong LLM, such as GPT-4, is employed. It is expected that this approach may surpass our method if video-to-text representations are further improved in the future. However, rather than solely focusing on performance, our framework prioritizes providing an interpretable perspective for VLMs in commonsense QA, giving users clear insights into the model's beliefs and reasoning paths.

Method	Model (Reasoner)	NExT-QA		IntentQA		VideoMME			
		Temporal	Causal	Temporal	Causal	Temporal	Spatial	Action	Object
VideoAgent	GPT-4 (1.8T)	64.5	72.7	64.1	66.5	-	-	-	-
VideoTree	GPT-4 (1.8T)	**67.0**	**75.2**	61.9	66.1	55.7	54.3	**54.2**	**52.6**
LLoVi	GPT-4 (1.8T)	61.0	69.5	65.5	**68.7**	52.2	**55.3**	51.8	50.8
Ours	VideoLLAVA (7B)	64.8	68.3	**66.1**	66.4	**55.9**	53.8	54.0	50.8

Table 10. Comparison results with state-of-the-art. Results for NExT-QA, IntentQA, and VideoMME are reported under its original test set. The 'Reasoner" in these approaches is similar to the "Prover" in our framework. The captioner for all methods is CogAgent. Despite other methods relying on a much stronger reasoning model, our approach yields competitive performance and reaches state-of-the-art results in four out of eight data partitions. Moreover, the reasoner we adopted is $250\times$ smaller than the others.

Model		Env-QA			
		State	Event	Order	Avg
Image-based VLMs	BLIP-2	30.6	28.8	40.2	33.2
	+Ours	39.5	34.5	45.8	39.9 (+6.7)
	LLaVA-1.5	31.3	30.7	42.8	34.9
	+Ours	40.5	36.1	46.2	40.9 (+6.0)
Video-based VLMs	VideoChat2	61.7	49.8	60.5	57.3
	+Ours	63.9	55.1	62.8	60.6 (+3.3)
	VideoLLaVA	60.5	50.4	61.0	57.3
	+Ours	63.3	55.5	63.2	60.7 (+3.4)

Table 11. Results on Env-QA. Incorporating our framework brings consistent improvement across the video- and image-based VLMs.

Metric	Model	NExT-QA	
		Original	Rewritten
Visual-text	CLIP	58.7	52.9
Text-text	LLaMA-3-8B	58.8	52.7
LLM-score	LLaMA-3-8B	59.7	54.3
Ours	LLaMA-3-8B	60.5	55.4

Table 12. Design of anchor frame localization. Our localization LLM outperforms competitive baselines. LLMs overall show a strong ability to retrieve relevant frames.

Modality		Video-text	Text	
Prv()		VideoLLaVA-7B	LLaMA-3-8B	GPT-4
NExT-QA	Original	60.5	57.1	59.6
	Rewritten	55.4	53.0	54.2

Table 13. Modality for proving entailment. Text-only reasoning paradigm achieves comparable performance only when a much stronger and larger ($250\times$) LLM, such as GPT-4, is employed.

Efficiency analysis of dynamic tree generation. To further validate the necessity of a dynamic strategy in entailment tree generation, we compare the efficiency of static and dynamic entailment tree approaches in Tab. 14. The results show that the number of LLM calls increases rapidly as the tree depth expands, introducing large time overheads.

	Static (Depth=)				Dynamic
	2	3	4	5	
Avg LLM calls	1	3	7	15	5.6
Acc (NExT-QA*)	52.0	53.4	55.6	55.3	55.4

Table 14. The efficiency comparison between static and dynamic entailment tree generation. 'Avg LLM Calls' is the average number of LLM calls made per statement during entailment generation. * indicates the de-biased set. By adopting the dynamic generation strategy, efficiency can be significantly improved without compromising performance.

Method	General VLM		VQA approaches			
	VideoChat2	VideoLlaVA	VideoAgent	VideoTree	LLoVi	Ours
Inf time(s)	7.5	6.2	51.0	34.6	40.3	38.2
Avg acc	49.0	50.8	61.6	60.9	58.6	**62.6**
Reasoner	VideoChat2 (7B)	VideoLlaVA (7B)	GPT-4 (1.8T)	GPT-4 (1.8T)	GPT-4 (1.8T)	VideoLlaVA (**7B**)

Table 15. Efficiency comparison. The average inference time for each video in the NExT-QA dataset is reported. VideoChat2 and VideoLlaVA are tested using 16 uniformly sampled frames (224×224) per video. For VideoAgent, VideoTree, and LLoVi, we adhered to their standard post-processing protocols for inference, whereas GPT-4 API served as the reasoning model.

By adopting the dynamic generation strategy, efficiency can be significantly improved as unnecessary decompositions will be pruned without compromising performance.

Efficiency analysis of overall framework Tab. 15 presents a comparative analysis of the accuracy-efficiency trade-off between our framework and existing general video-based VLMs, as well as state-of-the-art VQA methods. For this evaluation, we measured the average inference time per video on the NExT-QA dataset using NVIDIA-A600 GPUs. Specifically, VideoChat2 and VideoLlaVA were tested using 16 uniformly sampled frames (224×224) per video. For VideoAgent, VideoTree, and LLoVi, we adhered to their standard post-processing protocols for inference, whereas GPT-4 API served as the reasoning model. It can be seen that we achieve the best accuracy compared to other methods while maintaining a competitive inference speed of 38.2s (faster than VideoAgent and LLovi) and high parameter efficiency ($257\times$ fewer of the core reasoner than GPT-4 reasoners). This parameter efficiency further emphasizes

the practicality of our solution.

10. Qualitative results

Examples from the de-biased set. Fig. 6 showcases examples of Q&A pairs from the NExT-QA dataset before and after the de-biasing process. The original Q&A often exhibits textual biases or shortcuts between questions and options, which can be effectively mitigated through answer-set rewriting. The de-biased Q&A pairs compel VLMs to thoroughly comprehend both the video and text content to arrive at their answers. Therefore, this de-biasing procedure allows a more accurate evaluation of the VLMs' true commonsense reasoning abilities.

Entailment tree reasoning. In Fig. 7, we visualize the Q&A reasoning process through our proposed framework. Specifically, given the Q&A pair, we present the entire generated entailment tree and corresponding confidence scores for each statement produced during reasoning. Moreover, the grounded visual evidence is also presented. Our framework provides an interpretable window into VLMs about how the given Q&A is conducted in both the visual and textual modality.

11. Additional implementation details

Dataset overview (1) **NExT-QA** contains 5440 videos with an average length of 44s and approximately 52K questions. NExT-QA contains 3 different question types: Temporal, Causal, and Descriptive. In our experiments, we focus on the commonsense reasoning questions: Temporal and Causal. (2) **IntentQA** contains 4,303 videos and 16K multiple-choice question-answer pairs focused on reasoning about people's intent in the video. We perform a zero-shot evaluation on the test set containing 2K questions. (3) **VideoMME** comprises 2,700 QA pairs across 900 videos. Videos are annotated with 12 types of questions, including 4 types specifically designed for commonsense reasoning: temporal reasoning, spatial reasoning, action reasoning, and object reasoning.

Prompt designs. We provide our detailed designs of LLM prompts for implementing different functionalities in our framework, namely:

- *Video captioning*: fact-conditioned frame captioning (Fig. 8)
- *Entailment tree generation*: declarative statement transformation (Fig. 9), statement decomposition (Fig. 10)
- *Visual evidence grounding*: fact statement extractor (Fig. 11), fact statement retrieval (Fig. 12), evidence navigation (Fig. 13)
- *Visual-text statement verification*: statement verification via VLMs (Fig. 14)

Interval of Grounded moment The grounded interval is determined by the anchor frame and direction navigation. For '*look behind*', it starts at the anchor frame and ends at the video's end, while '*look ahead*' starts at the video's beginning and ends at the anchor. For '*look around*', a fixed 8-frame interval centered on the anchor frame is mapped back to the original video timestamp. Given the interval, we uniformly re-sample frames within the interval for VLM input, typically 8 or 16 frames, depending on the VLM's requirement.

Computing resources. Experiments are conducted on 4 NVIDIA-A6000 GPU and Azure Cloud APIs (for OpenAI models). The minimal GPU memory requirement is 24GB.

Reference

[1] Difei Gao, Ruiping Wang, Ziyi Bai, Xilin Chen. Env-QA: A Video Question Answering Benchmark for Comprehensive Understanding of Dynamic Environments. IEEE/CVF international conference on computer vision. 2021

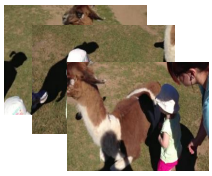
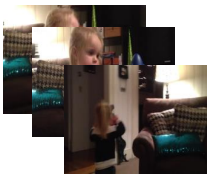
Video	Original Q&A	De-biased Q&A
	Q: Why does the woman caress the goat while the girl is staring at the goal? 1. Interested 2. show the kid it is ok 3. she is afraid 4. strolling 5. indicate to her to feed	Q: Why does the woman caress the goat while the girl is staring at the goal? 1. to calm the goat down 2. show the kid it is ok 3. show affection for the goat 4. teach the kid to interact gently 5. teach the kid about affection
	Q: How did the girl show excitement near the middle of the video? 1. pick up toy 2. put finger in mouth 3. standing 4. jumps 5. walking	Q: How did the girl show excitement near the mid of the video? 1. runs around 2. smiles 3. dances 4. jumps 5. claps hands
	Q: What did the man do when he approached the girl with the cake? 1. move the cake 2. bent down 3. help light candle 4. blow 5. excited and happy	Q: What did the man do when he approached the girl with the cake? 1. hug the girl 2. shake her hands down 3. help light candle 4. give a bouquet to the girl 5. Kiss the girl
	Q: What does the kid do after putting a finger into the bottle at the start? 1. reach his hand out 2. stick out tongue 3. touch white object 4. put bottle down 5. falls	Q: What does the kid do after putting a finger into the boittle at the start? 1. throw the bottle 2. take out the bottle 3. wash his hand 4. put bottle down 5. hold the bottle

Figure 6. Examples of original and de-biased Q&A, selected from NExT-QA dataset.

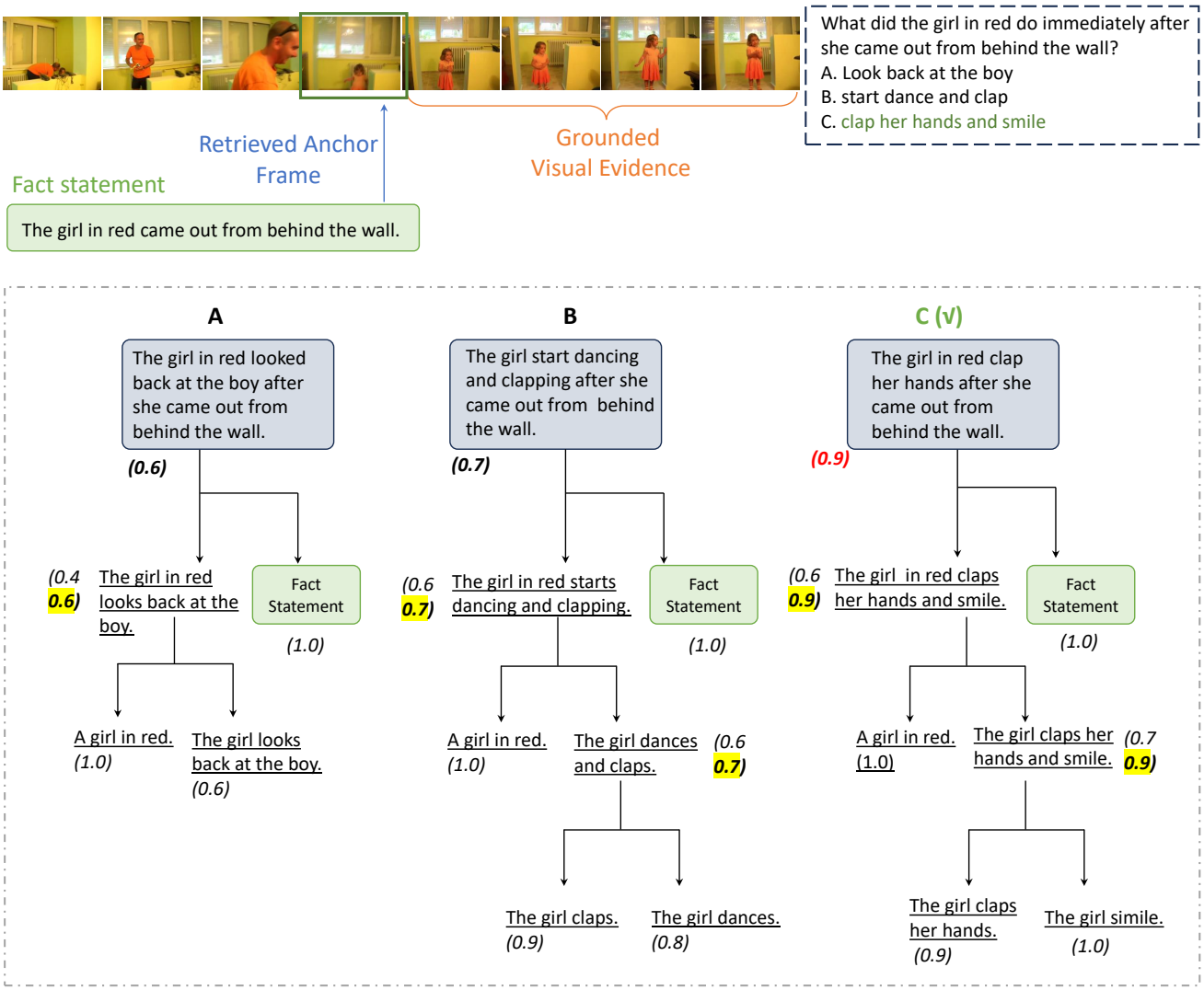


Figure 7. Examples of multi-choice QA inference of our framework. The highlighted confidence score indicates the proof score calculated from child statements.

User: Describe the given image, which represents the N-th frame in a video. Carefully analyze the video content, paying close attention to the objects, actions, and attributes of each object to provide a detailed description. Additionally, a fact statement related to a specific moment in the video is provided, which may offer cues about key objects or scenes to prioritize. You are also given the textual descriptions of previous frames in the video for reference.
Note: Do not just follow the fact statement, which is provided as a reference. You can only describe this image based on the image content and do not add any external knowledge to it.

Assistant:
<user_inputs_video >
Fact: <fact statement>
Previous descriptions before N-th frame: <captions before time N>

Figure 8. The prompt of fact-conditioned frame captioning for LLaVA-1.5.

User: You are presented with a question with corresponding multi-choice answer options. You are required to convert each option along with the question into a grammatical declarative statement sentence. Most importantly, make sure that proving the statement amounts to choosing that answer option over the other ones.
Note: do not modify the semantics of the sentence. Do not add new information or your own descriptions into the statements.

<Examples>:
Input:
Question: *Why does the brown cat watch the other cat eat food?*
(A). *Wants to go into box.*
(B). Wants to have a rest
(C). Waiting for his turn
(D). Playing with it
Output:
(A). The brown cat watch the other cat eat food because it wants to go into the box.
(B). The brown cat watch the other cat eat food because it wants to have a rest
(C). The brown cat watch the other cat eat food because it waits for his turn for food.
(D). The brown cat watch the other cat eat food because it's playing with it.

Assistant:
Input: <user_inputs>
Output:

Figure 9. The prompt of transferring Q&A into declarative statement for LLaMA-3.

User: Given a declarative statement, analyze the statement to extract distinct claims that could support this statement. Specifically, based on the claims, you need to decompose the statement into two shorter sub-statements, which can be utilized to verify the original statement jointly.
Note:
1. Each sub-statement should be verifiable and not overlap in content with the other one.
2. Make sure that the original statement is True if and only if both two sub-statements are True.
3. The sub-statement should be declarative sentences and avoid any hypothetical expression, such as "Let's assume", "consider whether".
4. If you think the given statement does not contain any verifiable facts, output "Decomposition failed: No worthy decomposition found."
5. Do not add additional information into the sub-statements that didn't indicate by the original statement.

<Examples>:
Input: *The man with spectacles looked to the camera after he looked down on the floor.*
Output:
(1) *The man with spectacles looked to the camera.*
(2) *The man with spectacles first looked down on the floor.*
Input: *The boy starts shake his legs to mimic the toy movement.*
Output:
(1) *The boy mimics the toy movement with his legs.*
(2) *The toy moves in shaking.*
Input: *The lady with jacket clapped her hands when the lady with microphone is performing.*
Output:
(1) *The lady with jacket clapped her hands.*
(2) *The lady with microphone is performing.*

Assistant:
Input: <user_inputs>
Output:

Figure 10. The prompt of statement decomposition for LLaMA-3.

User: Given multiple possible statements, your task is to extract a common fact claim. A fact claim is a statement that is acknowledged by all provided statements. Do not include any additional knowledge or information beyond what is explicitly present in the statements.

<Examples>:
Input:
(A). The brown cat watch the other cat eat food because it wants to go into the box.
(B). The brown cat watch the other cat eat food because it wants to have a rest
(C). The brown cat watch the other cat eat food because it waits for his turn for food.
(D). The brown cat watch the other cat eat food because it's playing with it.
Output*:*
The brown cat watch the other cat eat food.

Assistant:
Input: <user_inputs>
Output:

Figure 11. The prompt of fact statement extraction for LLaMA-3.

User: You are acting as a retriever. Given a query along with its structured semantic representation, your task is to identify the single most relevant frame from the provided semantic representations of all video frames. Carefully analyze the critical objects, actions, and attributes indicated by the query, compare them with all the candidate frames, and select the frame where the query is most likely to be represented.
Note: do not refuse to provide an answer and directly return the retrieved frame ID without any additional explanations.

<Examples>:
Input:
Query:
The boy in yellow is crawling out of the green mat.
<boy, in, yellow>, <boy, crawl, mat>, <boy, out of, mat>, <mat, in, green>
Candidate frames:
(1) *<boy, in, yellow>, <boy, pick, toy>*
(2) *<boy, in, yellow>, <boy, stand, _>, <boy, in front of, chair>*
(3) *<boy, play, toy>, <boy in yellow>*
(4) *<boy, on, mat>, <boy, sit, _>, <boy in yellow>,*
(5) *<boy, in, yellow,>, <boy, playing, _>, <boy, in, room>*
(6) *<boy, sit, mat>, <boy, in, room>*
Output frame ID:
(4)

Assistant:
Input: <user_inputs>
Output frame ID:

Figure 12. The prompt of retrieving fact statement for LLaMA-3.

User: You are acting as a navigator over the temporal dimension of a video. You will be presented with a question, a keyframe timestamp, and a fact statement describing an event or action occurring at that moment. Starting from this timestamp, your role is to determine the next direction to explore in the video, aiming to locate the segment most likely to answer the question. To guide your navigation, consider the semantic context of the entire video and prioritize the reasoning cues in the question (e.g., "what," "how," "why") and temporal indicators (e.g., "after," "while", "at the end of video") to make an informed decision about the next steps.
Note: you need to return your navigation from the following options:
(a) Look back
(b) Look behind
(c) Look around

<Examples>:
Input:
Question: *What does the boy do before crawling out of the green mat in the middle?*
Information of frames:
(1) *<boy, in, yellow>, <boy, pick, toy>*
(2) *<boy, in, yellow>, <boy, stand, _>, <boy, in front of, chair>*
(3) *<boy, play, toy>, <boy in yellow>*
(4) *<boy, on, mat>, <boy, sit, _>, <boy in yellow>,*
(5) *<boy, in, yellow,>, <boy, playing, _>, <boy, in, room>*
(6) *<boy, sit, mat>, <boy, in, room>*
Key frame timestamp and corresponding statement:
(4)
The boy is crawling out of the green mat.
Output navigation:
(a) Look back

Assistant:
Input: <user_inputs>
Output navigation:

Figure 13. The prompt of evidence navigation for LLaMA-3.

User: Are the following statements TRUE or FALSE in this video? Carefully watch the video content, paying close attention to the objects, actions, and attributes of each object in the video. For each statement, determine whether it is TRUE or FALSE in the video. Provide a response of 'TRUE' if the statement is correct, or 'FALSE' if the statement is incorrect.
Note: Apart from the video content, you cannot use additional information or rely on commonsense knowledge. Directly output 'TRUE' or 'FALSE' without adding explanations or any markers.

Assistant:
Input: <user_inputs_video > <user_inputs_text>
Output:

Figure 14. The prompt of statement verification for VideoLLaVA.