VITED : Video Temporal Evidence Distillation

Yujie Lu[1,2] Yale Song[2] William Wang[1] Lorenzo Torresani[2] Tushar Nagarajan[2]
[1]UC Santa Barbara [2]FAIR, Meta

Abstract

We investigate complex video question answering via chain-of-evidence *reasoning — identifying sequences of temporal spans from multiple relevant parts of the video, together with visual evidence within them. Existing models struggle with multi-step reasoning as they uniformly sample a fixed number of frames, which can miss critical evidence distributed nonuniformly throughout the video. Moreover, they lack the ability to temporally localize such evidence in the broader context of the full video, which is required for answering complex questions. We propose a framework to enhance existing VideoQA datasets with evidence reasoning chains, automatically constructed by searching for optimal intervals of interest in the video with supporting evidence, that maximizes the likelihood of answering a given question. We train our model (*VITED*) to generate these evidence chains directly, enabling it to both localize evidence windows as well as perform multi-step reasoning across them in long-form video content. We show the value of our* evidence-distilled *models on a suite of long video QA benchmarks where we outperform state-of-the-art approaches that lack evidence reasoning capabilities.*

1. Introduction

Video Question Answering (VideoQA) is a critical task towards general-purpose video understanding, for which large vision-language models (VLMs) [16, 44, 55] have recently demonstrated strong performance on a variety of benchmarks [17, 30, 32, 47–49]. Despite their popularity, these models typically excel at questions for which information is readily available throughout the video (e.g., high-level actions, colors/counts of objects, etc.), but struggle on long-form videos [9, 26, 30, 32] and on questions that require gathering and aggregating evidence from the full video [49]. For instance, answering a question like, "Why does the baby put their hand in their mouth at the beginning of the video?" requires more than just locating the action — it involves an intricate approach that gathers contextual clues both before and after the action to infer the underlying reason. For instance, examining what occurred just before or after the baby put their hand in their mouth and observing how they performed the action (e.g., slowly or quickly) can offer valuable insights into their intent.

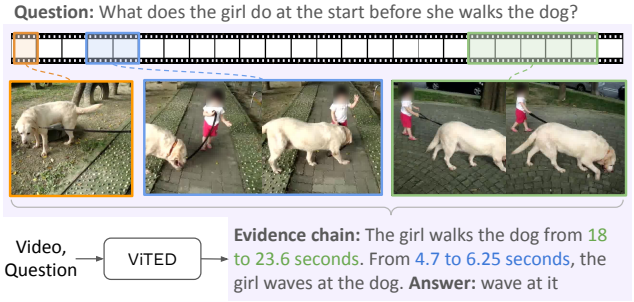

Figure 1. **Main idea.** We produce multiple, temporally localized pieces of evidence (the "evidence chain") to support complex reasoning in VideoQA. Our VITED model is trained to generate this evidence chain to enable temporally-grounded chain-of-thought reasoning in video.

Current models face limitations in effectively processing the temporal relationships among visual frames, as well as in bridging the gap between a question and the evidence needed to answer it. On the one hand, current models do not temporally ground their responses, and rather sample a fixed number of video frames at regular intervals, potentially missing important moments from the video (e.g., failing to identify the short window when the girl waves at dog in Fig. 1, left). Consequently, these models struggle to answer questions requiring complex *temporal understanding*, e.g., "what did the girl do right before interacting with the dog?" in a video where she approaches, picks up a ball, then engages with the dog; and *multi-hop reasoning*, e.g., "why did the baby put his hand in his mouth?" in a video where the mother feeds him with a spoon, followed by his uncomfortable expression and an attempt to remove the food.

On the other hand, *temporal grounding* models can indeed localize text queries to specific intervals in a video [13, 22, 33], but they require the text in advance — i.e., they can localize the evidence needed to answer the question if already provided with the evidence text, but they cannot identify and ground this evidence based on the question alone. For example, it might identify the moment when the dog approaches the girl (Fig. 1), but it may not establish how

this action relates to the question ("why did she wave?").

To address these limitations, we propose to consolidate evidence generation, temporal grounding and question answering into one model that we call **Vi**deo **T**emporal **E**vidence **D**istillation for Video Understanding (ViTED). Our model is a temporally-aware VLM that is trained to generate both answers to given questions, as well as temporally grounded *evidence chains* to support the answer — time intervals of the video together with textual clues within the temporal span (see Fig. 1, bottom). Since this kind of evidence data is not readily available, we present a framework to automatically synthesize high-quality evidence chains on top of existing VideoQA datasets. Specifically, we generate a pool of evidence containing textual evidence relevant to the question, extracted from video segments of various lengths and at multiple levels of detail (e.g., short clips with single actions, to high level activities across the full video). We then present a search-and-refinement algorithm over this evidence pool to find optimal sequences of evidence that are the most predictive of the correct answer. Finally, we augment the original VideoQA training data with our generated evidence chains, and train our model to predict both the answer and the evidence chain that supports it, thereby distilling the ability to localize and generate temporal evidence into the VLM.

We demonstrate the effectiveness of ViTED on 6 representative VideoQA benchmarks, CinePile [32], PerceptionTest [30], NExT-QA [48], STAR [47], MVBench [17]. We show that ViTED is on par with or outperforms state-of-the-art models trained with 10× more video instruction data. Additionally, we show that ViTED provides the most faithful and interpretable temporal evidence chain for the answer compared with existing VideoLLMs through human studies. Finally, ViTED achieves SOTA zero-shot performance on NExT-GQA [49] — a benchmark squarely focused on temporally grounded VideoQA — surpassing GPT-4 driven agent approaches, highlighting our generalizable evidence grounding capability.

In summary, we propose an approach that integrates evidence generation, grounding and reasoning towards complex video understanding. Our main contributions are:

- We propose a novel framework to generate and search for evidence chain-of-thought data from existing VideoQA datasets.
- We propose an *evidence distillation* approach to train a temporally-aware video model on our high-quality evidence data.
- Our ViTED sets new SOTA results among models of the same size on four VideoQA benchmarks, and surpasses GPT-4 driven agent on NExT-GQA, while providing high-quality explanations for its predictions.

2. Related Work

2.1. Video Understanding with LLMs

Recent LLM-based video models [16, 21, 24, 44, 55, 58] excel in video QA but lack temporal sensitivity, often missing key moments due to uniform frame sampling and struggling with multi-step reasoning. While recent methods [13, 33, 45] introduce time-aware representations for video-text grounding, they do not evaluate on general VideoQA. In contrast, we consolidate evidence generation, grounding, and question answering into a single model.

Agent-based or tool-assisted VLMs rely on retrieval [35, 43], memory [5, 7, 37], or modular reasoning [27, 53] for gathering video tokens. Our approach streamlines evidence generation and localization into a single-pass model, bypassing the need for multiple modules or API calls while enhancing performance.

2.2. Chain-of-Thought Reasoning in Videos

Chain-of-thought (CoT) [46] has been widely used to enhance the multi-step reasoning ability of LLMs. Various strategies have been proposed to ensure valid and logical reasoning paths through problem decomposition [60], deliberate search [51], and majority voting [41, 42]. Some works use knowledge distillation to enhance smaller models with the reasoning ability of larger models [18, 25, 36], or internalize this reasoning through implicit distillation [3].

In VLMs, CoT-based techniques [23, 59] improve reasoning by generating textual rationales or synthesizing multimodal infillings [34]. However, CoT for video understanding is under-explored. VIP [10] enhances CoT in LLMs and VLMs for video prediction, while methods like VSOR-CoT [39] improve video saliency prediction by reasoning about salient objects. MotionEpic [6] decomposes video tasks for better question answering using a Video-of-Thought framework. To the best of our knowledge, we are the first to *distill* chain-of-thought capabilities in VLMs, specifically for video understanding.

2.3. Visual Evidence

In image understanding, prior work has explored gathering visual evidence in the form of region of interest [52], interface layout [31] and programs [12]. In videos, prior work largely focuses on *frame-sampling* that varies sampling rate of the video depending on the content [38], employs keyframe extraction pipelines either through off-the-shelf approaches [14, 24] or based on learned text-frame similarity [19, 53]. Approaches for highlight detection cast visual evidence as a single time-window that matches a text description [28, 33, 37]. In contrast to the above, we propose to treat visual evidence in videos differently, as a series of temporally grounded descriptions that chain together to entail the answer to a question.

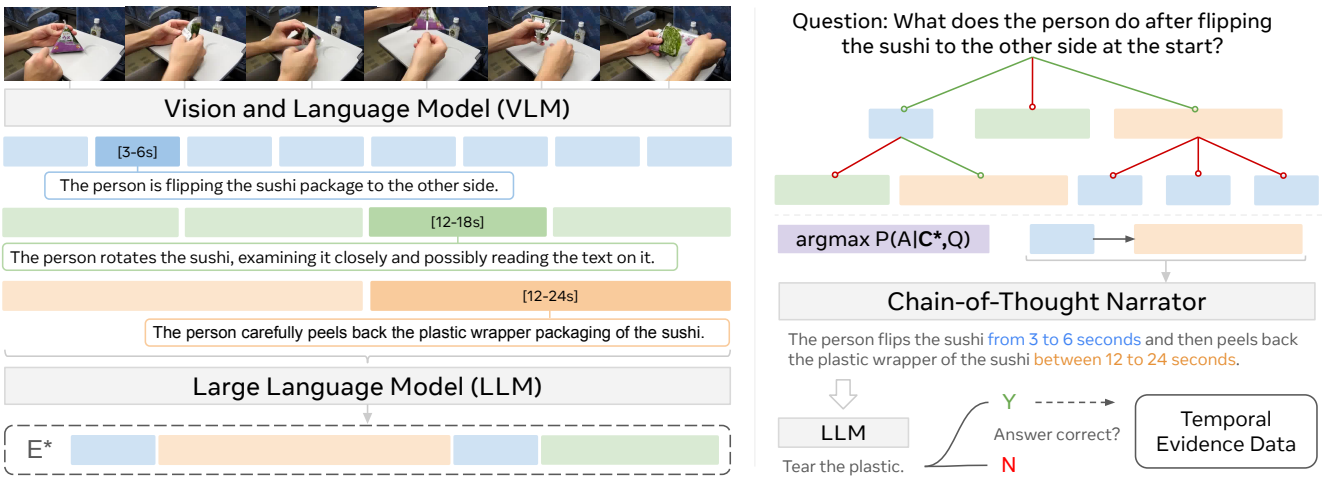

Figure 2. **Overview of ViTED evidence generation framework.** There are three main stages: (1) We first generate the evidence pool — detailed captions for segments at multiple granularities — and rank them based on relevance to the question (left, Sec.3.1) (2) Next, we search over the evidence pool to derive evidence chains that are most predictive of the target answer, and summarize it into a coherent and logical chain-of-thought (top-right, Sec. 3.2) (3) Finally, if the evidence chain successfully leads to the correct answer, we add it to our dataset for training our model (bottom-right, Section 3.3)

3. Approach

Our goal is to enable evidence-based video reasoning in VLMs by generating and distilling evidence chain data (i.e., time intervals with textual clues that support question-answering) into a model. This approach is crucial for datasets with complex reasoning questions. For example, in a study on NExT-QA, we found 54% of questions required localizing and reasoning over one or more salient windows (see Appendix A7). In short, we convert the standard question answering process (Q→A) into an evidence-based reasoning one (Q→Evidence + A).

Existing VideoQA datasets provide only question/answer text (e.g., Q: what happened after the dog barked? A: he jumped) — they do not provide temporal spans, or may provide a single, coarse span without explanation, making them insufficient for chain-of-thought evidence grounding. Temporal grounding datasets contain temporal spans for a query (e.g., Query: "A barking dog", Window: "4-9 seconds"), directly expressing the evidence to localize. However, questions in VideoQA datasets are more complex and do not reveal the evidence needed to determine the answer. We therefore propose a framework to construct evidence chains for videos. See Fig. 1 (right).

Our approach works as follows. First, we generate a pool of potential evidence from segmented video intervals using off-the-shelf VLMs (Sec. 3.1). Next, we filter and search through the evidence pool to identify the *most plausible* evidence chain using a combination of LLMs and VLMs (Sec. 3.2). Finally, we distill this reasoning into our temporally-aware VLM model by training a model to generate the evidence chains (Sec. 3.3).

3.1. Generating the Evidence Pool

We begin by generating a pool of potential evidence E for videos split into multiple segments and across a hierarchy of temporal granularities. Uniform segmentation or sparse sampling might miss key details because evidence may be unevenly distributed across a video, occurring at different granularities (e.g., static scenes vs. rapid actions). To address this, we propose a non-uniform segmentation of the video across N hierarchical levels. In each level, we create a sequence of sub-clips, each of length L, and separated by stride S between them. In our setup, we use $N = 5$, $(L, S) \in \{(1/16, 1/16), (1/8, 1/16), (1/4, 1/8), (1/2, 1/4), (1, 1)\}$.[1] This approach results in a set of video segments $\{vc_1, \ldots, vc_n\}$, covering five levels of granularity from global (full video context with $L = 1$ and $S = 1$) to fine-grained (small, localized segments with $L = 1/16$ and $S = 1/16$). See Fig. 2 (left) and Appendix A8 for details of the hierarchy.

We construct our evidence pool by prompting a VLM (LLaMA-3.2-Vision-Instruct-11B [4]) to generate evidence for each variable-length segment vc_i. In short, the model is asked to "*describe the contents of this segment that are relevant to the given question (but do not simply answer the question)*." The full prompt is in Appendix A9.

Fig. 2 (left) illustrates our evidence pool generation pipeline. After this process, we are left with a pool of candidate evidence $E = \{ev_1, ..., ev_m\}$ from five hierarchical levels that captures various temporal granularities of events in the full video, where each $ev_i = (t_s, t_e, \epsilon)_i$ representing the start time, end time, and evidence text for the vi-

[1] Both L and S are expressed as fractions of the full video duration.

sual chunk vc_i, respectively. We include an example of an evidence pool generated across all granularities in Appendix A8.

3.2. Refining and Searching for Evidence Chains

Each evidence piece $ev_i \in E$ may provide only partial information needed to answer the question and may lack explicit connections, such as temporal or cause-effect relationships among evidences. To construct a coherent evidence chain from this large, noisy pool, we propose a novel evidence search algorithm based on a text-only LLM (LLaMA-3.1-8B-Instruct [4]). This algorithm first narrows down the hypothesis space for possible evidence chains, then applies a beam search to identify the strongest chain.

Evidence Refinement To reduce noise, we begin by narrowing down the hierarchical evidence pool E to a reduced candidate pool E^* by ranking candidates directly using the LLM. We provide the full evidence pool to the LLM, and prompt it to *"Provide the evidence that will help reach the answer in a step-by-step manner. Limit your evidence chain to at most K steps."* This pool consists of K evidence segments representing a smaller, more manageable set of segments likely to be relevant to the question Q. The top-W evidence segments that are most likely to decode the correct answer are then selected from this reduced pool. This initial refinement narrows the search space and provides a focused foundation for constructing evidence chains. See Figure 2 (bottom left).

Evidence Chain Search Next, we search over sequences of evidence to identify high-likelihood chains, looking for the most coherent answer paths via iterative beam search. Starting from a refined initial beam of evidence segments, we initialize a beam with width $W = K/2$, half the size of the refined evidence pool. In each iteration, new evidence segments are appended to existing chains within the beam, generating expanded chains that may improve the likelihood of reaching the correct answer. Each expanded chain is then scored with the LLM with its likelihood of supporting the correct answer recalculated. Chains that meet a specified probability threshold T are retained as potential candidates, and the beam is updated to only include the top-W evidence chains based on likelihood scores. The evidence search process is summarized in Algorithm 1.

This process continues until either an evidence chain exceeds the threshold probability T or a fixed number of iterations is reached. The algorithm ultimately outputs the evidence chain C^* with the highest likelihood of correctly answering the question, ensuring a well-supported and coherent response. See Figure 2 (top right).

Evidence Chain Summarization and Filtering While the evidence chains C^* contain rich information related to the question, each piece of evidence was originally generated independently of each other, lacking event sequence information when simply concatenated. To address this, we summarize the entire evidence chain to be sequence-aware with a natural flow using the same text-based LLM, given its inherent ability to do chain-of-thought reasoning.

Algorithm 1 Evidence Chain Search

Input: Question Q, Answer A, Evidence Pool $E^* = \{ev_1, \ldots, ev_m\}$, Beam Width W, Threshold T
Output: Optimal Evidence Chain C^*
Initialize: Evidence chain $C \leftarrow \emptyset$, Beam $B \leftarrow \{\}$
Beam Search
Initialize beam $B \leftarrow \{ev_i : \text{top-}W \text{ by } P(A|Q, ev_i)\}$
while any C_i in B is updated **do**
  for each chain $C_i \in B$ **do**
    Expand C_i by adding $ev_j \in E^* \setminus C_i$
    Compute $P(A|Q, C_i \oplus ev_j)$
    If $P(A|Q, C_i \oplus ev_j) > T$, update C_i
  end for
  Update $B \leftarrow \{C_i \in B : \text{top-}W \text{ by } P(A|Q, C_i)\}$
end while
Set $C^* \leftarrow \arg\max_{C_i \in B} P(A|Q, C_i)$

Specifically, the LLM consolidates the interval (t_s, t_e), objects, events, and question cues in each evidence segment into a logical reasoning path that explicitly references the time intervals, and derives the final answer through step-by-step reasoning. In short, we prompt the LLM to "*convert this relevant evidence and its temporal span into a chain-of-thought reasoning based on the video.*" See Figure 2 (bottom left). The detailed prompt is in Appendix A9.

To ensure high-quality reasoning paths, we filter chains based on the LLM's ability to reach the correct answer. Specifically, we retain only those chains C_i^* that allow the LLM to correctly derive the answer A. Formally, we define a filtering criterion for a chain C_i^* as:

$$f(C_i^*) = \begin{cases} 1, & \text{if } \arg\max_{\hat{A}} \log p(\hat{A}|Q, C_i^*) = A, \\ 0, & \text{otherwise,} \end{cases} \quad (1)$$

where Q is the question, A is the correct answer, and $\log p(\hat{A}|Q, C_i^*)$ represents the log-likelihood of candidate answer $\hat{A}$ given Q and C_i^*. This ensures that only evidence chains that allow the model to logically connect the question, multi-hop evidence, and answer are retained. Successful chains are then added to the training data for the video model. This process yields a final dataset consisting of a video, question, evidence chain-of-thought, and answer.

3.3. Distilling Evidence Chains Into a Single Model

Finally, we distill the evidence chain information into our VITED model through curriculum training. For this, we

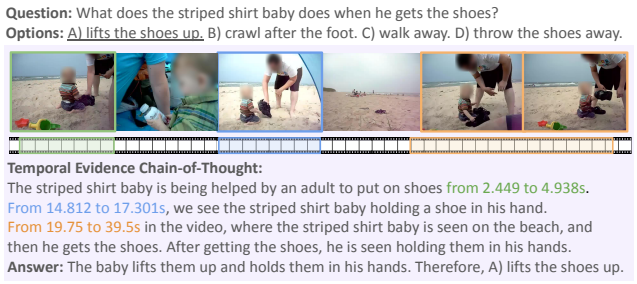

Figure 3. Example of temporal evidence on NExT-QA.

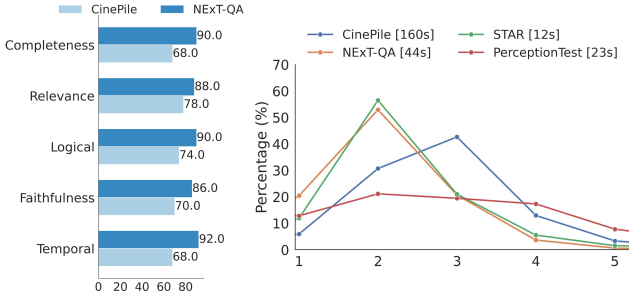

Figure 4. **Analysis of evidence quality. Left:** Human evaluation score on the quality of temporal evidence chain-of-thought. **Right:** Distribution of the number of hops in synthesized evidence chain across four datasets.

add an extra training stage, temporal evidence distillation, to the traditional instruction tuning in VLM models [33, 58].

Starting with a base VLM model, we perform instruction tuning (predicting answer tokens) as Stage-1. This enables the model to answer questions but without supporting evidence. Traditional chain-of-thought techniques (e.g., prompting the model to "think step-by-step") may be applied with limited success, as shown in our experiments. To improve this, we introduce *evidence distillation* in Stage-2, where the model learns to predict both the evidence chain (Sec. 3.2) and answer tokens, enabling it to reason across video segments. In both stages, the model is trained using *next token prediction* with cross-entropy loss to maximize the likelihood of generating evidence and/or answer tokens, conditioned on the video and question.

During inference, the model is required to generate the temporal evidence chain followed by the answer or directly answer the question depending on the input prompt.

In our experiments, we build on top of strong backbone VLM models, namely TimeChat [33] — a temporally-aware model suited to our fine-grained hierarchical temporal evidence data and LLaVA-Video [58] — a recent, SOTA VideoQA model.

4. Experiments

We evaluate models on a suite of VideoQA benchmarks, including CinePile [32] for long-video understanding, PerceptionTest [30] for low-level video perception, NExT-QA [48], STAR [47] and MVBench [17] for complex reasoning, and NExT-GQA [49] for temporal evidence grounding. For all datasets, we report MCQ accuracy.

Baselines We compare with state-of-the-art VLMs from existing literature as well as several variants of our approach to show the benefit of our temporal evidence distillation.

- **Base VLM** is the off-the-shelf VLM (TimeChat [33] and LLava-Video [58]) used for our experiments.
- **Chain-of-thought** implements the standard CoT mechanism by adding "let's think step-by-step" to the inference prompt. This is to show the non-trivial nature of grounding and generating temporal evidence in current VLMs.
- **Video Instruction Finetuning** is the model after standard instruction tuning on VideoQA datasets.
- **Dense Caption Distillation** is our model variant trained to generate dense captions (instead of evidence chains) by the same models that generate the evidence pool. This is to assess the value of temporal evidence beyond naive caption augmentation.
- **Temporal Evidence Distillation** is our proposed approach in Sec. 3.3 that trains a model to generate temporal evidence chains alongside direct answers, identifying sequences of temporal spans from multiple relevant parts of the video, together with visual evidence within them.

Note that all baselines are trained on the *same data* to allow apples-to-apples comparisons. For completeness, we also compare against state-of-the-art approaches from the literature that benefit from training on larger scale datasets.

Implementation and Training Details We use LLaMA-3.2-Vision-Instruct-11B as the evidence pool generator and LLaMA-3.1-8B-Instruct for evidence refinement and search in Sec. 3.1 and 3.2 respectively. We set beam width $W = 4$, threshold $T = 0.7$, and the maximum iterations as 3 in Algorithm 1. We use checkpoints from the official code repository as initializations for our TimeChat [33] and LLaVA-Video [58] base models.

We train VITED-TimeChat for 10 epochs, and VITED-LLaVA-Video for 1 epoch at each stage. We include dataset usage at each training stage and evaluation dataset details in Appendix A11. VITED is trained on 291K video question answering samples from public datasets such as NExT-QA, STAR, and PerceptionTest. Details of full data mix and additional hyperparameters can be found in Appendix A11. We use a batch size of 64, with 8 nodes of 8-V100 (32G) machine. For TimeChat-based models, we unfreeze only the image Q-Former, video Q-Former, and linear layer and process 96 input frames for each video, while for LLava-Video, we unfreeze only the adapter and the LLM backbone. We use LoRA [11] with a rank 32 for TimeChat-based and rank 128 for LLaVA-based VITED.

Models	Params	CinePile	PercepTest	NExT-QA	STAR	MVBench	NExT-GQA
State-of-the-art Video LLMs							
SeViLA [53]	4B	-	-	73.8	64.9	-	16.6
VideoLLaVA [20]	7B	22.51	-	-	-	-	-
LongVA [56]	7B	-	-	68.3	-	-	-
LLaMA-3.2V [4]	11B	39.55	52.65	67.58	45.62	44.72	11.64
InternVideo2 [44]	6B	-	63.4	78.6	-	67.2	-
LLaVA-OneVision [16]	7B	46.42	57.1	79.4	66.24	55.91	0
TimeChat Base							
TimeChat [33]	7B	13.67	20.34	21.71	21.03	17.08	0.00
TimeChat (Chain-of-Thought)	7B	14.27	20.96	25.60	17.54	22.05	0.00
TimeChat (Video Instruction Tuning)	7B	55.86	57.39	71.05	68.39	47.48	0.00
ViTED (Dense Caption Distillation)	7B	57.85	59.94	71.94	69.99	50.12	0.00
ViTED (Temporal Evidence Distillation)	7B	**58.98**	**63.66**	**73.42**	**70.99**	**50.46**	**27.61**
LLaVA-Video Base							
LLaVA-Video [58]	7B	53.77	**67.90**	83.20	66.88	58.60	0.04
LLaVA-Video (Chain-of-Thought)	7B	54.82	66.91	**84.61**	65.34	59.50	0.03
LLaVA-Video (Video Instruction Tuning)	7B	54.97	67.03	83.93	67.74	59.55	0.00
ViTED (Dense Caption Distillation)	7B	54.79	67.07	83.91	67.56	58.53	0.00
ViTED (Temporal Evidence Distillation)	7B	**56.39**	67.50	84.13	**68.23**	**60.95**	**25.19**

Table 1. **VideoQA benchmark results.** Our temporal evidence distillation is consistently better than dense caption distillation, direct video instruction tuning, and naive chain-of-thought. ViTED built on LLaVA-Video achieves SOTA on 4 out of 6 VideoQA benchmarks.

Model	LLM	NExT-GQA	
		IoP@0.5	Acc@GQA
LLaMA-3.2V [16]	LLaMA-3 [4]	19.8	11.6
FrozenBiLM [50]	DeBERTa [8]	23.7	17.5
SeViLA [53]	Flan-T5 [2]	22.9	16.6
LLoVi [54]	GPT-4 [29]	38.0	26.8
ViTED	LLaMA-2 [40]	**41.2**	**27.6**

Table 2. **Results on NExT-GQA [49]**. IoP@0.5 and Acc@GQA represent intersection over prediction of evidence and accuracy of grounded question answering.

E-pool	S-CoT	NExT-QA			
		Temporal	Causal	Descriptive	Avg.
✗	✗	68.22	71.65	74.87	71.05
✓	✗	73.50	74.35	78.87	74.79
✗	✓	69.91	73.76	81.98	73.80
✓	✓	73.46	76.34	81.03	76.14

Table 3. **Ablation of evidence data** with and without Evidence pool (E-pool) or chain-of-thought summarization (S-CoT) stages.

4.1. Quality of Generated Evidence

To begin, we analyze the quality of evidence chains themselves. First, we verify whether generated evidence chains capture sufficient detail to answer questions (i.e., without re-watching the video). We do this by prompting a text-only LLaMA-3.1-8B-Instruct with a question and an evidence chain, and measuring the accuracy of selecting the correct answer on NeXT-QA. Through this process, our generated evidence chain yields 72.05%, significantly improving over the conventional chain-of-thought (51.71%). The remaining gap can be attributed to shortcomings in the base VLM responsible for generating the evidence pool. These issues include hallucinations in video segment descriptions and vague descriptions that omit crucial information, which in turn propagate errors throughout the entire evidence chain.

Next, we evaluate the quality of evidence directly by asking two human annotators to score a subset of evidence chains across five key aspects on a 3-point scale (good, average, bad). These are (1) Temporal: Does the temporal window match the evidence text? (2) Faithfulness: Is the evidence faithful to the video content? (3) Logical: Is the reasoning logical across evidence? (4) Relevance: How relevant is the evidence chain to the video/question? and (5) Completeness: Does the evidence chain capture all required

Stage-1	Stage-2	CinePile						STAR					NExT-GQA	
		CRD	NPA	STA	TEMP	TH	Avg.	Int.	Seq.	Pre.	Fea.	Avg.	IoP	Acc
✗	✗	13.04	11.86	17.16	11.86	1.92	13.62	20.02	21.31	24.84	19.18	21.04	0.00	0.00
✓	✗	57.05	54.15	58.99	44.59	63.46	55.86	63.30	70.20	74.68	72.04	68.39	0.00	0.00
✗	✓	59.59	57.20	59.81	48.30	66.99	58.12	66.93	73.16	79.03	75.92	71.76	41.60	27.42
✓	✓	59.80	58.89	61.57	49.23	65.38	58.98	65.76	73.06	75.80	75.31	70.99	41.22	27.61

Table 4. **Stage-1 vs. Stage-2 training.** Effect of Stage-1 (standard video instruction tuning, Q→A) and Stage-2 (temporal evidence finetuning, Q→A, Q→E,A, Q→A,E) on performance (Sec. 3.3).

Model	NExT-QA			Avg
	Temporal	Causal	Descriptive	
No Evidence	68.22	71.65	74.87	71.05
Direct Multi-Evidence	72.65	75.25	81.21	75.34
GT-Guided Sampling	72.36	74.31	80.50	74.64
ViTED	73.46	76.34	81.03	76.14
w/o Hier	70.75	74.61	80.54	74.28
w/o Search	69.69	73.78	78.72	73.22
w/o Multi-Hop	70.35	74.92	83.25	74.74

Table 5. **Ablation on Evidence Data Framework.** 1) using off-the-shelf model to generate evidence chain, 2) hierarchical evidence pool (Hier), evidence chain search (Search), multi-hop temporal evidence (Multi-Hop).

information in the video to answer the question? Full instructions are in Appendix A12. Figure 4 (left) presents the average score in each category based on human annotations. We find that while it is harder to generate reliable evidence on datasets with longer videos, our approach still scores over 80.4% on average across the five aspects, indicating the effectiveness of our synthetic data pipeline.

Finally, we show statistics of our synthesized evidence data in Figure 4 (right). We find that, for benchmarks with longer videos, such as CinePile tend to favor larger number of hops compared with shorter duration video benchmarks, such as STAR and NExT-QA. We show a qualitative example of our temporal evidence in Figure 3. See Appendix A14 for more examples and analysis.

4.2. ViTED for Video Question Answering

In Table 1, our ViTED with temporal evidence distillation significantly outperforms video instruction tuning across seven video-based QA benchmarks, achieving gains such as $+3.12\%$ on CinePile, $+6.27\%$ on PerceptionTest, and $+27.61\%$ on NExT-GQA with TimeChat-base. Additionally, ViTED-LLaVA-Video surpasses the SOTA baseline LLaVA-Video by $+2.62\%$ on CinePile, $+2.35\%$ on STAR, and $+25.19\%$ on NExT-GQA. Our temporal evidence distillation also consistently surpasses video instruction tuning and dense caption distillation on top of our base video models. This result underscores the model's strength in handling temporally distributed evidence, which is critical for accurately interpreting video content where events unfold over time. While LLaVA-Video excels in video question answering and TimeChat in video temporal grounding, neither is tailored for grounding the evidence entailed in questions, leading to near-zero accuracy on NExT-GQA evidence grounding task.

4.3. ViTED for Grounding Visual Evidence

From Table 1, ViTED shows the benefits of temporal evidence capability in achieving strong results on NExT-GQA. This is significant as evidence grounding arises naturally from our model's training, without relying on explicit grounding modules [1, 22] or manually labeled evidence grounding data [15, 49]. We further compare with the current SOTA in evidence grounding on NExT-GQA in Table 2 across two standard metrics, Intersection over Prediction (IoP) and Accuracy (Acc) of grounded question answering (GQA). ViTED achieves a remarkable IoP@0.5 score of 41.22 and Acc@GQA of 27.61%, outperforming even GPT-4 driven agent approaches like LLoVi [54], and making it the best model to date.

4.4. Ablation Experiments

We present experiments ablating key design choices in our approach: the importance of different modules in evidence generation (Table 3), the role of training stages (Table 4), various evidence generation strategies (Table 5). Additional ablations on evidence-pool generators and hyperparameters are in Appendix A13. Unless otherwise specified, we show single-task fine-tuning results on NExT-QA, using the TimeChat backbone for all the ablations.

Can we directly generate evidence with VLMs? First, we explore simpler methods of evidence generation — by simply prompting off-the-shelf VLMs to directly generate them. We test two strategies: (1) Direct Multi-Evidence Grounding, where we prompt the model to "*provide a detailed sequence of information to help answer the question in the form [start time, end time] {supporting evidence}*", given the question and options; and (2) GT-Guided Evidence Sampling, where we sample up to three evidence

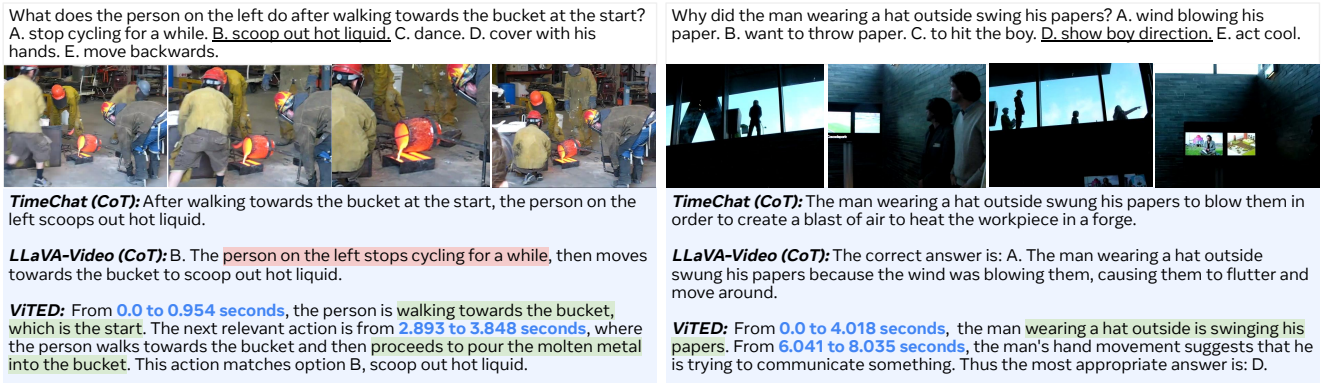

Figure 5. **Examples of generated evidence chains.** Compared to traditional chain-of-thought approaches, ViTED demonstrates temporal evidence generation and reasoning capabilities, accurately analyzing the sequence of actions in the video to reach the correct final answer. Colored text and highlights are for visualization only and correspond to wrong evidence, correct evidence and temporal localization windows of generated evidence (**blue text**).

chains from the same model, and use the ground-truth answer to select the most appropriate one. Full details of each strategy are in Appendix A10.

Our results in Table 5 (top) show that both methods are able to produce coherent evidence chains that enhance performance compared to the model with no additional evidence. Both baseline strategies surpass the baseline without evidence by a large margin, which indicates the importance of evidence. The strategy with the ground truth guidance is even worse, indicating that the input guidance with answer can lead model to make up non-plausible evidence chains. Meanwhile, we can see our ViTED surpasses both strategies, indicating that our evidence generation pipeline is effective. Overall, our approach outperforms these direct variants highlighting the need for more elaborate evidence generation strategies.

In Table 5 (bottom), we show additional ablations examining the necessity of: hierarchical evidence pools by utilizing only a single level (S=1, L=1) of the hierarchy (w/o Hier); evidence search by directly using the filtered evidence pool E^* (w/o Search); and multi-hop evidence by forcing evidence to be single hop (w/o MultiHop). We observe that there is smaller or no performance degradation in "Descriptive" type of questions, compared with "Temporal" and "Causal" types. This indicates the design of our hierarchical evidence pool, evidence search, and multi-hop are essential to complex video understanding.

Importance of evidence generation stages We ablate different stages of the evidence generation pipeline, namely the need for the evidence pool itself (Sec. 3.1), the chain-of-thought summarization of evidence (Sec. 3.2) and their interplay. We simply replace the evidence pool with a single generated evidence chain (similar to the direct approaches above) and/or drop the summarization step from the pipeline, leaving evidence chains in their raw form.

Our results in Table 3 show that both stages are important to achieve optimal distillation. Although, without the evidence pool, our ViTED can still achieve on par results as our full model on 'Descriptive' category, it is far worse in the 'Temporal' and 'Causal' category. While when the summarization module is dropped, the model performance degrades severely in the 'Descriptive' category, and slightly on 'Causal' category. Without these two evidence generation stages, the model is consistently worse than our full ViTED across three aspects.

Are both training Stage-1 and 2 essential? In Table 4, we ablate Stage-1 and Stage-2 in our proposed temporal evidence curriculum training strategy. We show accuracy on subcategories: Character and Relationship Dynamics (CRD), Narrative and Plot Analysis (NPA), Setting and Technical Analysis (STA), Temporal (TEMP), Theme Exploration (TH) on CinePile, Interaction (Int.), Sequence (Seq.), Prediction (Pre.), and Feasibility (Fea.) on STAR. Without Stage-1 and Stage-2 is equivalent to our base TimeChat model. Adding video instruction tuning (Stage-1) leads to significant improvements on CinePile and STAR, but no gain on NExT-GQA which requires evidence grounding. With our Stage-2 (temporal evidence finetuning), we achieve SOTA results on CinePile, STAR and NExT-GQA.

5. Conclusion

We proposed a novel pipeline to synthesize high-quality chain-of-evidence data on top of existing video understanding data, and a video model to distill this temporal video evidence data via curriculum training. Our results show notable improvements over state-of-the-art models with larger sizes and more training data, by unlocking temporally-grounded chain-of-thought reasoning in videos.

6. Acknowledgements

We sincerely thank Shraman Pramanick for his insightful brainstorming and thoughtful feedback on our early manuscript.

References

[1] Qirui Chen, Shangzhe Di, and Weidi Xie. Grounded multi-hop videoqa in long-form egocentric videos, 2024. 7

[2] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, and et al. Siddhartha Brahma. Scaling instruction-finetuned language models, 2022. 6

[3] Yuntian Deng, Yejin Choi, and Stuart Shieber. From explicit cot to implicit cot: Learning to internalize cot step by step, 2024. 2

[4] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, and et al. Angela Fan. The llama 3 herd of models, 2024. 3, 4, 6, 18

[5] Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. Videoagent: A memory-augmented multimodal agent for video understanding, 2024. 2

[6] Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong Li Lee, and Wynne Hsu. Video-of-thought: Step-by-step video reasoning from perception to cognition. In *International Conference on Machine Learning*, 2024. 2

[7] Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. Ma-lmm: Memory-augmented large multimodal model for long-term video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13504–13514, 2024. 2

[8] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention, 2021. 6

[9] Xuehai He, Weixi Feng, Kaizhi Zheng, Yujie Lu, Wanrong Zhu, Jiachen Li, Yue Fan, Jianfeng Wang, Linjie Li, Zhengyuan Yang, Kevin Lin, William Yang Wang, Lijuan Wang, and Xin Eric Wang. Mmworld: Towards multi-discipline multi-faceted world model evaluation in videos, 2024. 1

[10] Vaishnavi Himakunthala, Andy Ouyang, Daniel Rose, Ryan He, Alex Mei, Yujie Lu, Chinmay Sonar, Michael Saxon, and William Yang Wang. Let's think frame by frame with vip: A video infilling and prediction dataset for evaluating video chain-of-thought, 2023. 2

[11] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. 5

[12] Yushi Hu, Otilia Stretcu, Chun-Ta Lu, Krishnamurthy Viswanathan, Kenji Hata, Enming Luo, Ranjay Krishna, and Ariel Fuxman. Visual program distillation: Distilling tools and programmatic reasoning into vision-language models, 2024. 2

[13] De-An Huang, Shijia Liao, Subhashree Radhakrishnan, Hongxu Yin, Pavlo Molchanov, Zhiding Yu, and Jan Kautz. Lita: Language instructed temporal-localization assistant, 2024. 1, 2

[14] KeplerLab. Katna: Tool for automating video keyframe extraction, video compression, image autocrop and smart image resize tasks. https://github.com/keplerlab/katna, 2019. 2

[15] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. TVQA: Localized, compositional video question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1369–1379, Brussels, Belgium, 2018. Association for Computational Linguistics. 7

[16] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer, 2024. 1, 2, 6, 18

[17] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. Mvbench: A comprehensive multi-modal video understanding benchmark, 2024. 1, 2, 5, 16

[18] Liunian Harold Li, Jack Hessel, Youngjae Yu, Xiang Ren, Kai-Wei Chang, and Yejin Choi. Symbolic chain-of-thought distillation: Small models can also "think" step-by-step, 2024. 2

[19] Hao Liang, Jiapeng Li, Tianyi Bai, Xijie Huang, Linzhuang Sun, Zhengren Wang, Conghui He, Bin Cui, Chong Chen, and Wentao Zhang. Keyvideollm: Towards large-scale video keyframe selection, 2024. 2

[20] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection, 2024. 6

[21] Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models, 2024. 2

[22] Kevin Qinghong Lin, Pengchuan Zhang, Joya Chen, Shraman Pramanick, Difei Gao, Alex Jinpeng Wang, Rui Yan, and Mike Zheng Shou. Univtg: Towards unified video-language temporal grounding, 2023. 1, 7

[23] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022. 2

[24] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models, 2024. 2

[25] Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. Teaching small language models to reason. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1773–1781, Toronto, Canada, 2023. Association for Computational Linguistics. 2

[26] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding, 2023. 1

[27] Juhong Min, Shyamal Buch, Arsha Nagrani, Minsu Cho, and Cordelia Schmid. Morevqa: Exploring modular reasoning models for video question answering, 2024. 2

[28] WonJun Moon, Sangeek Hyun, SangUk Park, Dongchan Park, and Jae-Pil Heo. Query-dependent video representation for moment retrieval and highlight detection, 2023. 2

[29] OpenAI. Gpt-4 technical report, 2024. 6

[30] Viorica Pătrăucean, Lucas Smaira, Ankush Gupta, Adrià Recasens Continente, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Joseph Heyward, Mateusz Malinowski, Yi Yang, Carl Doersch, Tatiana Matejovicova, Yury Sulsky, Antoine Miech, Alex Frechette, Hanna Klimczak, Raphael Koster, Junlin Zhang, Stephanie Winkler, Yusuf Aytar, Simon Osindero, Dima Damen, Andrew Zisserman, and João Carreira. Perception test: A diagnostic benchmark for multimodal video models, 2023. 1, 2, 5, 16

[31] Yijun Qian, Yujie Lu, Alexander G. Hauptmann, and Oriana Riva. Visual grounding for user interfaces. In *2024 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2024) - Industry Track*, 2024. In preparation. 2

[32] Ruchit Rawal, Khalid Saifullah, Miquel Farré, Ronen Basri, David Jacobs, Gowthami Somepalli, and Tom Goldstein. Cinepile: A long video question answering dataset and benchmark, 2024. 1, 2, 5, 16

[33] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding, 2024. 1, 2, 5, 6

[34] Daniel Rose, Vaishnavi Himakunthala, Andy Ouyang, Ryan He, Alex Mei, Yujie Lu, Michael Saxon, Chinmay Sonar, Diba Mirza, and William Yang Wang. Visual chain of thought: Bridging logical gaps with multimodal infillings, 2024. 2

[35] Chuyi Shang, Amos You, Sanjay Subramanian, Trevor Darrell, and Roei Herzig. Traveler: A multi-lmm agent framework for video question-answering. *ArXiv*, abs/2404.01476, 2024. 2

[36] Kumar Shridhar, Alessandro Stolfo, and Mrinmaya Sachan. Distilling reasoning capabilities into smaller language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7059–7073, Toronto, Canada, 2023. Association for Computational Linguistics. 2

[37] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, Yan Lu, Jenq-Neng Hwang, and Gaoang Wang. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18221–18232, 2024. 2

[38] Reuben Tan, Ximeng Sun, Ping Hu, Jui-Hsien Wang, Hanieh Deilamsalehy, Bryan A. Plummer, Bryan Russell, and Kate Saenko. Koala: Key frame-conditioned long video-llm. 2024. 2

[39] Yunlong Tang, Gen Zhan, Li Yang, Yiting Liao, and Chenliang Xu. Cardiff: Video salient object ranking chain of thought reasoning for saliency prediction with diffusion, 2024. 2

[40] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, and et al. Shruti Bhosale. Llama 2: Open foundation and fine-tuned chat models, 2023. 6

[41] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Huai hsin Chi, and Denny Zhou. Rationale-augmented ensembles in language models. *ArXiv*, abs/2207.00747, 2022. 2

[42] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models, 2023. 2

[43] Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent, 2024. 2

[44] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Chenting Wang, Guo Chen, Baoqi Pei, Ziang Yan, Rongkun Zheng, Jilan Xu, Zun Wang, Yansong Shi, Tianxiang Jiang, Songze Li, Hongjie Zhang, Yifei Huang, Yu Qiao, Yali Wang, and Limin Wang. Internvideo2: Scaling foundation models for multimodal video understanding, 2024. 1, 2, 6

[45] Yueqian Wang, Xiaojun Meng, Jianxin Liang, Yuxuan Wang, Qun Liu, and Dongyan Zhao. Hawkeye: Training video-text llms for grounding text in videos, 2024. 2

[46] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. 2

[47] Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. STAR: A benchmark for situated reasoning in real-world videos. In *Thirty-fifth Conference on Neural Information Processing Systems (NeurIPS)*, 2021. 1, 2, 5, 16

[48] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9777–9786, 2021. 2, 5, 16

[49] Junbin Xiao, Angela Yao, Yicong Li, and Tat Seng Chua. Can i trust your answer? visually grounded video question answering, 2024. 1, 2, 5, 6, 7, 16

[50] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Zero-shot video question answering via frozen bidirectional language models, 2022. 6

[51] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: deliberate problem solving with large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2024. Curran Associates Inc. 2

[52] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity, 2023. 2

[53] Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. Self-chained image-language model for video localization and question answering, 2023. 2, 6

[54] Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. A simple llm framework for long-range video question-answering, 2024. 6, 7

[55] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding, 2023. 1, 2

[56] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024. 6

[57] Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, 2024. 18

[58] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data, 2024. 2, 5, 6

[59] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models, 2024. 2

[60] Denny Zhou, Nathanael Scharli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Olivier Bousquet, Quoc Le, and Ed Huai hsin Chi. Least-to-most prompting enables complex reasoning in large language models. *ArXiv*, abs/2205.10625, 2022. 2