Enhancing Video-LLM Reasoning via Agent-of-Thoughts Distillation

Yudi Shi[1,2], Shangzhe Di[1,2], Qirui Chen[1,2], Weidi Xie[1,†]
[1]School of Artificial Intelligence, Shanghai Jiao Tong University, China
[2]Coop. Medianet Innovation Center, Shanghai Jiao Tong University, China
https://zhengrongz.github.io/AoTD/

Abstract

This paper tackles the problem of video question answering (VideoQA), a task that often requires multi-step reasoning and a profound understanding of spatial-temporal dynamics. While large video-language models perform well on benchmarks, they often lack explainability and spatial-temporal grounding. In this paper, we propose ***A****gent-****o****f-****T****houghts* ***D****istillation (****AoTD****), a method that enhances models by incorporating automatically generated Chain-of-Thoughts (CoTs) into the instruction-tuning process. Specifically, we leverage an agent-based system to decompose complex questions into sub-tasks, and address them with specialized vision models, the intermediate results are then treated as reasoning chains. We also introduce a verification mechanism using a large language model (LLM) to ensure the reliability of generated CoTs. Extensive experiments demonstrate that AoTD improves the performance on multiple-choice and open-ended benchmarks.*

1. Introduction

Video Question Answering (VideoQA) refers to a critical task that offers a natural interface for human-machine interaction through language [33, 42, 43, 49]. This synergy of visual content and language enhances the accessibility of AI systems for the general public, allowing users to query complex visual content with natural language. By encompassing tasks such as action recognition, object detection, and scene understanding, VideoQA serves as a comprehensive benchmark for evaluating AI's ability to interpret videos, addressing the fundamental questions of 'who', 'what', 'when', and 'where' that are crucial to understand daily life activities, pushing the boundaries of what AI systems can interpret from dynamic visual content.

Recent literature has primarily explored two avenues in VideoQA. The first involves training large video language models (Video-LLMs) through direct instruction-tuning, using videos paired solely with corresponding questions and answers [1, 22, 24, 53]. While these models excel on public benchmarks, they often lack explainability and struggle with spatial-temporal grounding. This limitation hinders their ability to provide clear reasoning, which is essential for real-world applications where transparency and interpretability are critical [29].

Conversely, an emerging approach utilizes agent-based systems that decompose complex questions into manageable sub-tasks, each addressed by specialized tools [15, 17, 37]. The results are then aggregated to form a coherent answer. Theoretically, such approach naturally offers greater interpretability, as the reasoning process is divided into explainable steps that can be independently assessed. However, our experiments indicate that current video understanding tools are not strong enough for building reliable agent-based systems. In addition, the high memory demands and time-consuming nature of these systems present significant challenges for their practical use.

In this paper, we aim to leverage the advantage of both research lines, enhancing Video-LLM by integrating Chain-of-Thoughts (CoTs) into instruction-tuning, with the CoTs being constructed from the outputs of specialized agent models, capturing the step-by-step reasoning procedure, as illustrated in Figure 1.

In specific, we start by systematically evaluating the off-the-shelf models tailored for atomic video understanding tasks, such as action recognition [39, 41] and language grounding [23], using well-annotated datasets. This comprehensive evaluation allows us to pinpoint the most effective tools for each sub-task, thus laying a robust foundation for constructing reliable chains. Moreover, this process also provides a critical assessment of the broader capabilities of visual models across general and complex scenes, offering valuable insights for future research within the community.

In addition, we introduce a verification mechanism using a large language model (LLM), designed to assess if the generated CoTs adhere to a clear, step-by-step reason-

†: Corresponding author.

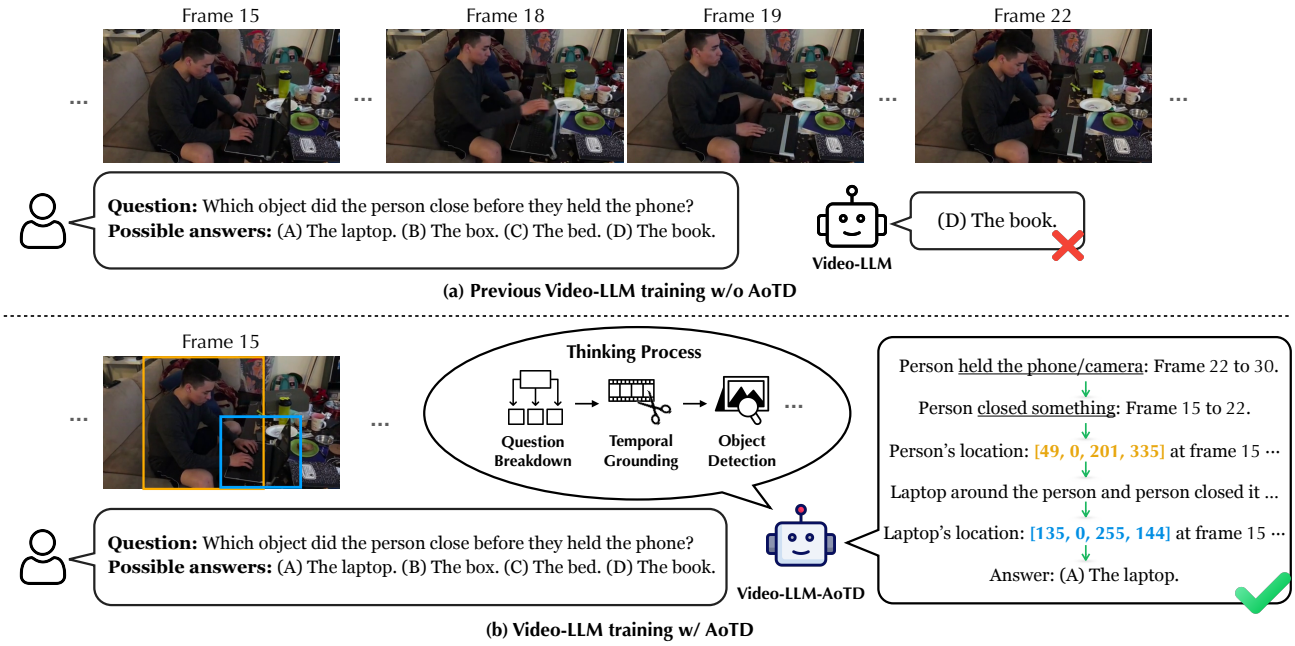

Figure 1. Our method, **AoTD**, distills multi-step reasoning and spatial-temporal understanding into a single generative video-language model. When addressing complex VideoQA tasks, the model trained with AoTD (as shown in (b)) enables to generate a step-by-step reasoning to get the correct answer. In contrast, previous models trained solely on question-answer pairs (as in (a)) generate only a final answer, often without intermediate reasoning, which can lead to incorrect conclusions.

ing process and incorporate essential information for answering the queries effectively. This mechanism filters out low-quality or logically inconsistent reasoning paths. The remaining CoTs that pass this verification are then distilled into large generative video-language models, significantly enhancing both their performance and interpretability, ultimately leading to the development of more robust, accurate, and interpretable VideoQA systems.

In summary, our contributions are three-fold: (i) we propose a novel approach for enhancing Video-LLMs by distilling high-quality CoTs into their instruction-tuning process. These CoTs capture step-by-step reasoning paths, improving both the model's performance and its interpretability; (ii) to automatically construct the CoTs for any dataset, we employ an agent-based system to decompose complex VideoQA questions into simpler sub-tasks, leveraging off-the-shelf vision models to handle each sub-task. The intermediate outputs from these models can therefore be collected as CoTs for addressing the corresponding visual question; (iii) through extensive experiments, we demonstrate that our distilled model outperforms existing methods across both multiple-choice and open-ended VideoQA benchmarks, enabling to deliver not only accurate answers but also comprehensive reasoning explanations.

2. Related Work

Video-language models (Video-LLMs). Recent works such as VideoLLaMA2 [4], LLaVA-NeXT-Video [54] and VideoChat2 [19], with their excellent architecture design and reasonable instruction-tuning data collection, have achieved a new level of zero-shot results in VideoQA task. However, current end-to-end models still lack interpretability for questions, as well as the ability to think and visually process complex problems in multiple steps, which is an important part for embodied learning and autonomous driving.

Visual Programming and Agents. With the progress of LLMs, some recent works [5, 15, 37, 45] begin to try to use LLM as planner to solve the complex reasoning task in real scenarios. They attempt to decompose the question into some easier sub-questions, and use different specialist models as agents to solve these sub-questions, and finally gather them to get the answer of the raw question. These models demonstrate a strong ability to obtain trustworthy answers based on the intermediate evidence they get, but they lag far behind the end-to-end model in terms of inference speed.

Visual Chain-of-Thoughts (CoTs). The potential of Chain-of-Thought (CoT) reasoning [40, 46] extends from NLP to the visual domain, highlighting a growing interest in applying this approach across various fields. Numerous studies have incorporated CoTs into visual understanding tasks [11, 30, 36, 55], utilizing powerful Multi-Modal Large Language Models (MLLMs) for generating CoTs or adopting tool-based architectures for sequential problem-solving.

Recent innovations, for example, Visual Program Distillation (VPD) [16] and Fact [10] attempt to address these

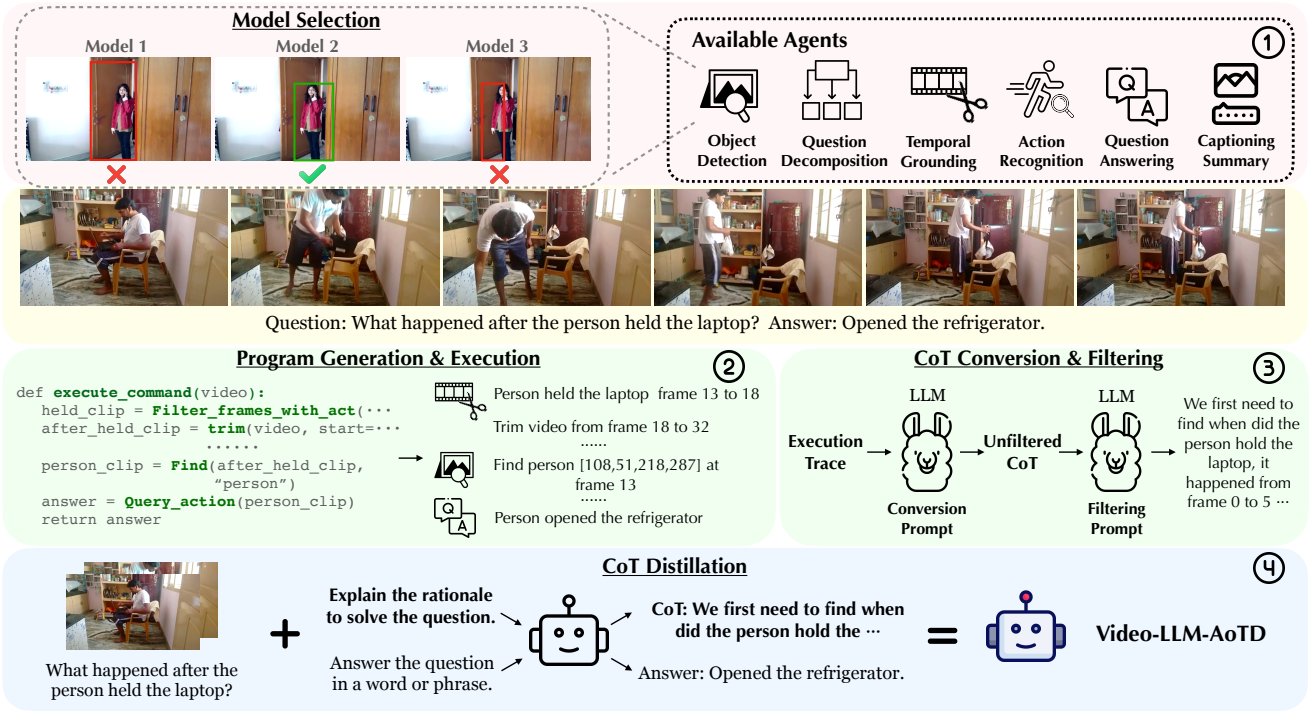

Figure 2. Overview on Agent-of-Thoughts Distillation (AoTD). **Step 1:** Selecting best-performing agents for each sub-task to construct an agent-based system. **Step 2:** Decomposing question into executable program and leveraging chosen models to solve it sequentially to generate execution trace. **Step 3:** The execution trace is converted and filtered by LLM to produce high quality natural language CoTs. **Step 4:** Distilling CoTs into Video-LLM with two forms of prompt, allowing it achieve a balance between concise answers and comprehensive rationales. The final model is Video-LLM-AoTD.

issues by maintaining the accuracy and diversity of CoTs, while leveraging MLLMs to generate them directly. These approaches decompose complex tasks into code programs, call upon expert models to handle sub-tasks, and utilize the resulting CoTs as training data to fine-tune visual-language models. This process significantly improves the models' ability to generate detailed rationales. Despite the progress in image understanding, there remains a notable oversight in video domains, where reasoning chains can be particularly effective due to the complex spatial-temporal dynamics of video understanding tasks. This is the focus of our paper.

Concurrent Work. In the recent literature, we notice two work that share similar idea with ours, specifically, Video-STaR [56] construct CoTs using videos and existing labels for instruction-tuning, yet they do not develop an agent-based system. Meanwhile, MotionEpic [8] introduces a Video-of-Thought reasoning framework that integrates video spatial-temporal scene graphs, marking a significant stride towards more nuanced video reasoning.

3. Agent-of-Thoughts Distillation

In this paper, we propose a novel approach, termed Agent-of-Thoughts Distillation (AoTD), to enhance the Video-LLMs by training them with multi-step Chain-of-Thoughts. Specifically, we start by developing an agent-based video understanding system, to generate multi-step reasoning chains that address complex video questions. These reasoning chains are then distilled into one Video-LLM through instruction-tuning. By combining the strengths of agent-based systems and large generative models, our proposed AoTD enables to build more reliable and interpretable VideoQA systems. Figure 2 illustrates the entire process of our method.

3.1. Problem Formulation

Given a video clip with t frames, $\mathcal{V} = \{x_1, \ldots, x_t\}$, and a set of n questions $\mathcal{Q} = \{q_1, q_2, ..., q_n\}$, our goal is to train a Video-LLM capable of producing both concise answers and comprehensive rationales. Depending on the suffix prompt p_s, the model can generate different types of outputs. The process can be formulated as:

$$\{a_i, \mathcal{S}_i\} = \Phi(\mathcal{V}, q_i, p_s), \ \mathcal{S}_i = \{\emptyset\} \text{ or } \{s_{i,1}, \ldots, s_{i,k}\}$$

where q_i denotes the i-th question, a_i is the answer in free-form text, and $\mathcal{S}_i$ represents the rationale, consisting of the reasoning process. If the prompt specifies to only generate the answer, $\mathcal{S}_i = \{\emptyset\}$. Otherwise, if the prompt requires the generation of rationales, $\mathcal{S}_i = \{s_{i,1}, \ldots, s_{i,k}\}$, where each $s_{i,j}$ corresponds to a reasoning step.

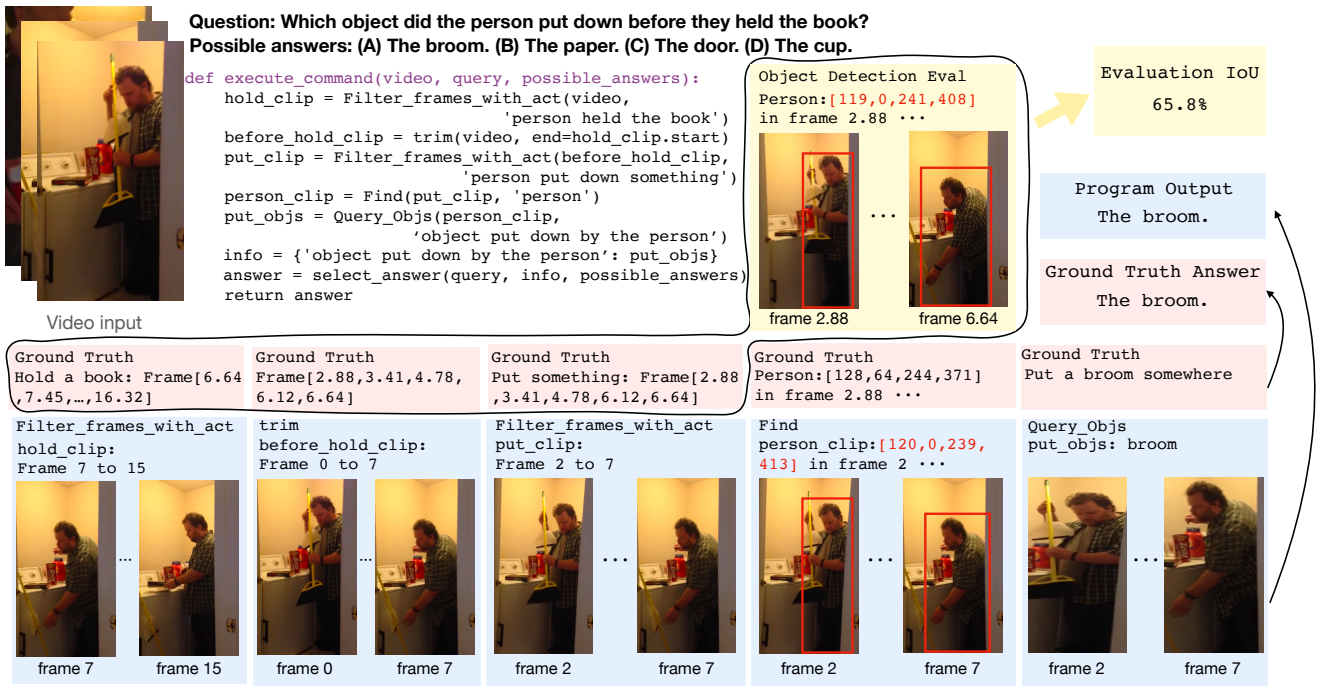

Figure 3. Program execution process in an agent-based system. We uniformly sample 32 frames from the video, and to ensure scale consistency, the frame ids of key frames are normalized into these 32 frames. The blue boxes represent the program execution steps, the red boxes denote the ground truth for each step. The combination of red and yellow boxes represents one example process of evaluating Object detection model candidates.

Discussion. Unlike existing models that are instruction-tuned on VideoQA datasets using simple question-answer pairs, which bypass the intermediate thought process, our approach emphasizes the importance of training with Chain-of-Thoughts (CoTs). In the following section, we outline the process for generating high-quality CoTs from existing VideoQA datasets.

3.2. CoT Construction with Agent-based System

Recent work, such as STAR [42], has introduced the idea of employing executable symbolic programs, to directly decompose questions into sub-tasks. When combined with scene graphs that contain comprehensive video information from key frames—such as object locations, interactions, and actions—these programs facilitate the generation of concise Chain-of-Thoughts (CoTs) through direct execution of symbolic operations. However, datasets of this nature are limited in scale, we therefore propose to first build an agent-based system, capable of breaking down complex questions into simpler sub-tasks, and the intermediate outputs from this system can then be employed to construct CoTs for any existing VideoQA dataset.

Agent-based VideoQA. Assuming we are given a video input ($\mathcal{V}$), questions ($\mathcal{Q}$), and a set of visual models ($\mathcal{M} = \{\phi_{\text{act}}, \phi_{\text{det}}, \ldots, \phi_{\text{qa}}\}$), an LLM-based agent core ($\pi(\cdot)$) processes the question along with the documentation of the visual models ($\mathcal{T}$), which includes variables and functionalities. The agent subsequently decomposes the question into sub-tasks formatted as Python code, and resolves them by invoking the appropriate visual models through function calls. It is important to note that the visual models can be arranged in various orders depending on the specific question, ensuring flexibility in problem-solving.

Specifically, as illustrated by the example in Figure 3, the

Sub-task	Model name	Metric	Number (%)
question decomposition	CodeQwen1.5-Chat (7B) [2]	Acc	52.7
	GPT-3.5-Turbo [31]		73.1
	DeepSeek-Coder-Instruct (6.7B) [6]		85.7
object detection	OWL-ViT v1 [26]	IoU	47.3
	GLIP [20]		58.9
	OWL-ViT v2 [28]		63.0
temporal grounding	LITA (13B) [18]	IoU / Recall	11.7 / 20.2
	TimeChat (7B) [35]		13.9 / 23.1
	UniVTG [23]		24.7 / 35.3
action recognition	InternVideo2 (1B) [39]	Top1-Acc	7.6
	Open-VCLIP [41]		8.9
	LLaVA-NeXT-Video-DPO (7B) [54]		18.2
question answering	LLaMA-VID (7B) [21]	Acc	43.5
	SeViLA [48]		46.5
	LLaVA-NeXT-Video-DPO (7B) [54]		53.4

Table 1. Sub-tasks definition and evaluation results. We choose 3 model candidates for each sub-task and evaluate them in STAR training set with the corresponding metrics. Models with best performance are placed at the bottom of each column.

question is first decomposed into a series of sub-tasks, including temporal grounding, object detection, and question answering. The corresponding specialized models are then executed sequentially to address these sub-tasks, ultimately yielding the final answer y_i:

$$\{\phi_{\text{ground}}, \phi_{\text{det}}, \phi_{\text{qa}}\} := \pi(q_i, \mathcal{T}),$$
$$y_i = \phi_{\text{ground}}(\mathcal{V}) \rightarrow \phi_{\text{det}}(\mathcal{V}) \rightarrow \phi_{\text{qa}}(\mathcal{V})$$

CoT Construction. To ensure the correctness of outputs at all the intermediate steps, we leverage the training set from STAR for hyperparameter tuning, enabling us to identify the most effective model for each sub-task within the agent-based system. By following the provided programs, we evaluate the performance of the corresponding vision models on tasks such as object detection and action recognition. Given the availability of complete reasoning chains, we independently assess each sub-task using ground truth data for all preceding steps.

As shown in Table 1, we present the evaluation results for the various sub-tasks. Specifically, for **question decomposition**, we compare several code LLMs, with DeepSeek-Coder-Instruct achieving the highest performance, outperforming even GPT-3.5-Turbo. In **object detection**, OWL-ViT v2 records the highest Intersection over Union (IoU) score, showcasing its superior open-vocabulary detection capability. The results for **temporal grounding** indicate that while UniVTG leads in performance, there remains a need for further advancements in this area. In **action recognition**, our evaluations show that generative models outperformed discriminative models, likely due to the fine-grained action list provided by the STAR dataset. However, the performance of both model types reveals significant room for improvement. Finally, in the **one-hop question answering** sub-task, all models perform admirably, with LLaVA-NeXT-Video-DPO standing out as a top performer, consistent with its strong results on other benchmarks.

With these high-performing models, we implement the agent-based approach on VideoQA datasets that consist solely of QA pairs. During the execution of the programs, we record all intermediate outputs to construct the CoTs. Since the outputs from these vision models vary in format—such as bounding boxes and free-form text—we employ another LLM to translate the execution trace into natural language for better use in the distillation process. Detailed examples are provided in Appendix C.

3.3. CoT Verification

To refine the quality of reasoning chains for VideoQA samples, we implement a two-step verification: (i) we filter execution traces to retain only those, where the program can reach correct output. For multiple-choice datasets, the output must match the correct answer exactly, while for open-ended datasets, we prompt the LLM to verify correctness, accounting for format differences; (ii) we prompt the LLM to evaluate the logical coherence and usefulness of the reasoning chains in solving the problem. The model assesses whether the CoTs follow a clear, step-by-step reasoning process and provides a binary evaluation ('Yes' or 'No') to indicate their quality (detailed prompts are included in Appendix D). This two-step approach ensures that only high-quality CoTs are retained for further distillation.

Dataset	Description	# Labels	# CoTs
AGQA	Compositional	25.0K	5.4K
ANetQA	Compositional	25.0K	3.6K
STAR	Compositional	45.7K	11.2K
NExT-QA	Temporal & Causal	34.1K	12.1K
CLEVRER	Spatial & Temporal	21.0K	-
EgoQA	Ego-centric	7.8K	-
Total		**158.6K**	**32.3K**

Table 2. Dataset statistics. The column "# Labels" indicates the number of VideoQA pairs, which include the video, query, possible answers (multiple-choice), and the correct answer. "# CoTs" refers to the number of CoTs generated using our agent-based system for each dataset.

In Table 2, we provide the statistics for the remaining generated CoTs for different datasets. We primarily select compositional QA datasets, as these require the model to process spatial-temporal information from different events comprehensively.

3.4. Step-by-step Distillation

In this section, we describe the process of distilling the generated CoTs into a Video-LLM. This distillation enhances the model's ability for spatial-temporal video understanding and multi-step reasoning, thereby improving its performance on complex VideoQA tasks.

In specific, using the generated CoTs, we can build the dataset $D = \{(\mathcal{V}_j, q_j, \hat{y}_j, c_j, p_s)\}_{j=1}^{N}$, where N is the total number of samples in the distilling dataset, $\mathcal{V}_j$ is the video input, q_j is the question, $\hat{y}_j$ is the ground-truth answer, c_j is the generated CoT, p_s is the task-specific suffix prompt, to distinguish different tasks, for example, for multiple-choice VQA, the prompt can be: "Answer with the option's letter from the given choices directly and only give the best option", and for open-ended VQA, the prompt can be: "Answer in one word or phrase". Please refer to detailed prompts in Appendix D.

At distillation stage, we minimize the cross-entropy loss of predicting both the answer and the CoTs, we replace the suffix prompt p_s with "Explain the rationale to answer the question", to indicate whether we want a question answer or a rationale to explain the thinking steps. Following [10]

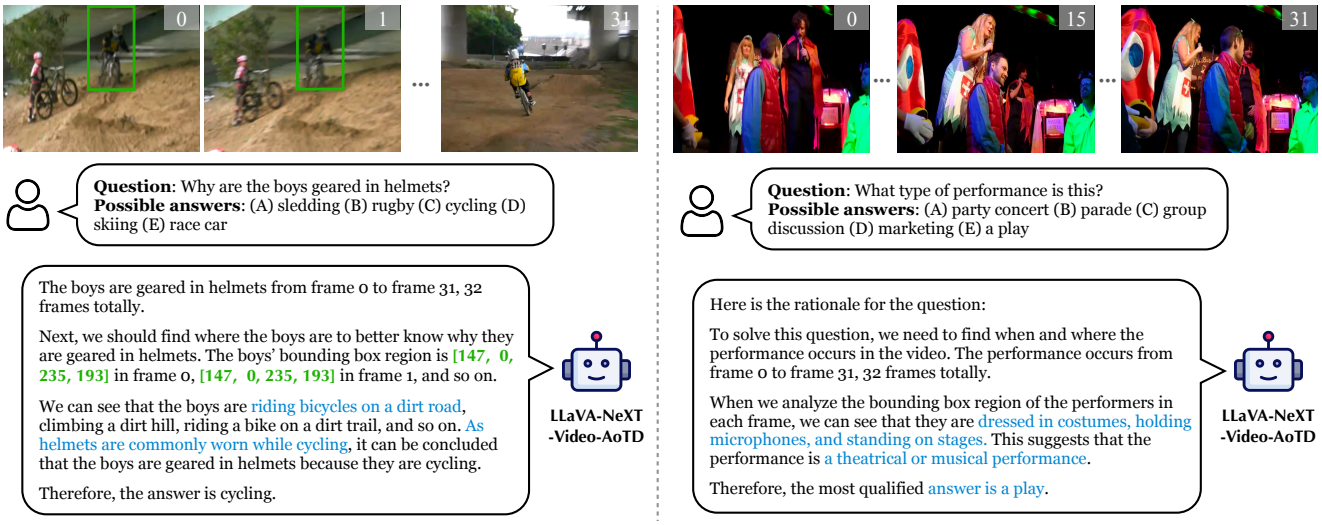

Figure 4. **Visualization of rationales.** LLaVA-NeXT-Video-AoTD can output rationales containing both spatial-temporal grounding of key information and step-by-step thinking process to solve the question.

and [16], our optimization objective is:

$$\begin{aligned}\mathcal{L} &= \mathcal{L}_{\text{label}} + \lambda \mathcal{L}_{\text{rationale}} \\ &= \sum_{j=1}^{N} \ell(\Phi(\mathcal{V}_j, q_j, p_s), \hat{y}_j) + \lambda \ell(\Phi(\mathcal{V}_j, q_j, p_s), c_j)\end{aligned}$$

Here, we set λ to 1 to ensure the importance of answer and rationale are equally considered. Notice that, not all the QA pairs can generate qualified CoT. In that case, the $\mathcal{L}_{\text{rationale}}$ will be set to 0.

Dataset	Size		Type	Train	Eval
	train	eval			
MC-VQA					
STAR [42]	45.7K	7.1K	Compositional	✓	✓
NExT-QA [43]	34.1K	5.0K	Temporal & Causal	✓	✓
CLEVRER [47]	21.0K	-	Spatial-temporal	✓	✗
Perception-Test [33]	-	11.5K	General	✗	✓
MVBench [19]	-	4.0K	General	✗	✓
VideoMME [9]	-	2.7K	General	✗	✓
VSIBench [44]	-	5.0K	spatial-temporal	✗	✓
OE-VQA					
AGQA [14]	25.0K	2.0K	Compositional	✓	✓
ANetQA [50]	25.0K	2.0K	Compositional	✓	✓
EgoQA [13]	7.8K	-	Ego-centric	✓	✗
Activitynet-QA [49]	-	8.0K	General	✗	✓
Video-ChatGPT [24]	-	3.0K	General	✗	✓

Table 3. Training and evaluation datasets statics.

4. Experiments

In this section, we present the experimental setup (Sec. 4.1) and results on various VideoQA benchmarks (Sec. 4.2). Extensive ablation studies have also been conducted to further examine the contributions of our approach in Sec. 4.3, and an evaluation on the quality of rationales generated by the distilled model is made in Sec. 4.4.

4.1. Experimental Setup

Base model. We use LLaVA-NeXT-Video (7B) [54] (LNV for short) as base Video-LLM, which has shown remarkable performance on image-centric tasks, for example image question answering [51]. We present comparison on naive instruction-tuning with video question answering dataset or with additional CoT distillation. For CoT conversion and verification, we prompt LLaMA-3.1-8B with the manually-designed instruction and some in-context examples. Detailed prompts are provided in Appendix D.

Instruction tuning. We utilize both multiple-choice and open-ended QA data, along with the generated CoTs, to fine-tune the base video question answering model, as summarised in Table 2. The resulting distilled model is named **LLaVA-NeXT-Video-AoTD** (LNV-AoTD for short). Additionally, as baseline, we also train another version of the model using only the basic QA data, which we refer to as **LLaVA-NeXT-Video-Instruct** (LNV-Instruct for short).

Evaluation benchmarks. We conduct extensive evaluations on Multiple-Choice Video QA (MC-VQA) and Open-Ended Video QA (OE-VQA). We report the top-1 accuracy for all MC benchmarks, which means the proportion of the output equal to the answer. We report a GPT-assessed Acc. and Score with the help of GPT-3.5-turbo-0613 for all OE benchmarks. For each question, GPT delivers a binary decision indicating whether the output is correct or incorrect, along with a similarity score reflecting the degree of alignment between the output and the correct answer. The term 'Acc.' refers to the percentage of correct outputs, while 'Score' represents the average similarity scores. For the

Model	MVBench (Acc.)	VideoMME (Acc.)	STAR (Acc.)	NExT-QA (Acc.)	Perception-Test (Acc.)	VSIBench (Acc.)
Proprietary Models						
Gemini 1.5 Pro [12]	-	75.0	-	-	-	45.4
GPT4-V [32]	43.7	59.9	-	-	-	-
Open-source Models						
LLaMA-VID (7B) [21]	41.9	25.9	-	-	44.6	-
Video-LLaVA (7B) [22]	41.0	39.9	-	-	44.3	-
VideoChat2 (7B) [19]	51.1	33.7	59.0*	68.6*	47.3	-
VideoLLaMA2 (7B) [4]	53.4	45.1	58.5*	62.3*	49.6	-
LLaVA-NeXT-Video (7B) [54]	46.5*	41.0*	52.4*	61.6*	47.5*	19.7*
LLaVA-NeXT-Video-Instruct (7B)	53.4	43.2	72.2	77.1	50.3	26.7
LLaVA-NeXT-Video-AoTD (7B)	55.6	45.0	74.3	77.6	50.6	28.8

Table 4. Comparison with Video-LLMs on MC-VQA benchmarks. LLaVA-NeXT-Video-AoTD outperforms all other baselines the and the version without CoT distillation. * means results reproduced by ourseleves. Results without signs are retrieved from [4] and [44].

evaluation on AGQA and ANetQA, due to the large volume of test set, we test on a subset of samples. We evenly select the benchmark in-domain and out-of-domain for testing to ensure a comprehensive and reasonable evaluation of the model capability. Noted that though VSIBench has both MC and OE questions, it doesn't need GPT for score, so we classify it into MC benchmarks for convenience. Detailed statistics for evaluation benchmarks are shown in Table 3.

4.2. Quantitative Results

We divide the comparison into two parts: the first focuses on comparing the distilled model with other baselines, while the second examines the difference between the instruct version and the AoTD version. Note that, the latter part will be mainly compared and discussed, to demonstrate the model's improvement relative to its previous performance, as well as establishing the transferability of the method across models.

MC-VQA performance. As shown in Table 4, our LLaVA-NeXT-Video-AoTD achieves superior performance across all benchmarks. Several key observations can be made: (i) comparing to the base model, even a simple instruction-tuning on certain VideoQA datasets significantly enhances the model's question-answering performance. This improvement is notable, as the base model was primarily trained on static images and struggled with video understanding; (ii) our model, instruction-tuned with CoT distillation, demonstrates further performance enhancements across all benchmarks, particularly on the compositional VideoQA benchmark (STAR) and comprehensive benchmarks (VideoMME, MVBench). This suggests that our AoTD method effectively improves the model's ability to address complex problems and interpret spatial-temporal scenes; (iii) the distilled model consistently outperforms all other baselines across almost all benchmarks, even when compared to more powerful models. This finding shows that our method effectively bridges performance gaps created by varying model components.

OE-VQA performance. As shown in Table 5, LLaVA-NeXT-Video-AoTD outperforms the Instruct variant across all open-ended VideoQA benchmarks. Notably, it achieves a greater percentage increase compared to the MC-VQA benchmarks, suggesting that CoT distillation may be more effective for open-ended generation than for multiple-choice selection. While the distilled model scores higher than most models listed in the table, it does not surpass LLaVA-NeXT-Video on certain benchmarks. We conjecture this is due to the model's extensive training on images, that can also benefit the question answering without requiring complex reasonings, as also suggested by the findings in VideoLLaMA2 [4]. Additionally, the inherent challenges of evaluating open-ended VQA may influence the results. Assessments conducted by GPT can be biased or inaccurate, and the metrics we employ primarily indicate general trends rather than providing absolute accuracy.

4.3. Ablation Study

Analysis on CoT filtering. To prove the effectiveness of our filtering mechanism, we trained an alternative model without CoT filtering while maintaining all other settings, *i.e.*, using 36.3K verified CoTs for distillation. As shown in Table 9, the model's performance declines significantly on both the Multiple-Choice VQA and Open-Ended VQA benchmarks when the CoT filtering mechanism is not utilized. This confirms that employing large language models (LLMs) to filter CoTs is crucial for enhancing data quality.

Analysis on model transferability. As AoTD is a distillation method that leverages Chain-of-Thoughts (CoTs), it can theoretically be applied to any Video-LLMs. To assess the transferability of our method, we conduct experiments on another very recent model, LLaVA-OneVision (7B) [3]. As shown in Table 9, our method also demonstrates significant improvements on the benchmarks, showing the transferability and robustness of the approach. Due to the rapid advancements in the computer vision field, evaluating all

Model	ANetQA (Acc./Score)	AGQA (Acc./Score)	Video-ChatGPT (Score) Corr.	Deta.	Cont.	Temp.	Cons.	ActivityNet (Acc./Score)
Proprietary Models								
Gemini 1.5 Pro [12]	-	-	-	-	-	-	-	56.7/-
GPT4-V [32]	-	-	4.09	3.88	4.37	3.94	4.02	59.5/-
Open-Source Models								
VideoLLaMA (7B) [53]	-	-	1.96	2.18	2.16	1.82	1.79	12.4/1.1
Video-ChatGPT (7B) [24]	-	-	2.50	2.57	2.69	2.16	2.20	35.2/2.7
LLaMA-VID (7B) [21]	-	-	2.96	3.00	3.53	2.46	2.51	47.4/3.3
Video-LLaVA (7B) [22]	-	-	2.87	2.94	3.44	2.45	2.51	45.3/3.3
VideoChat2 (7B) [19]	-	-	3.02	2.88	3.51	2.66	2.81	49.1/3.3
VideoLLaMA2 (7B) [4]	-	-	3.09	3.09	3.68	2.63	3.25	49.9/3.3
LLaVA-NeXT-Video (7B) [54]	46.4/3.3*	27.4/2.2*	3.26*	3.22*	3.77*	2.47*	2.99*	54.3/3.2*
LLaVA-NeXT-Video-Instruct (7B)	47.1/3.1	59.3/3.4	2.96	2.81	3.35	2.42	2.82	50.0/3.3
LLaVA-NeXT-Video-AoTD (7B)	53.9/3.4	60.9/3.6	3.11	3.00	3.60	2.41	2.91	53.2/3.4

Table 5. Comparison with Video-LLMs on OE-VQA benchmarks. LLaVA-NeXT-Video-AoTD improves performance in all open-ended benchmarks compared with the Instruct version. * means results reproduced by ourseleves. Results without signs are retrieved from [4].

models and benchmarks is prohibitively infeasible. Thus, we focus on assessing some representative models against selected benchmarks to provide a representative evaluation.

Model	Filtering	MVBench (Acc.)	STAR (Acc.)	AGQA (Acc. / Score)
LNV-AoTD	✗	53.7	73.3	59.5/3.5
LNV-AoTD	✓	55.6	74.3	60.9/3.6
Onevision-Instruct	-	59.2	75.8	65.6/3.7
Onevision-AoTD	✓	60.5	76.6	65.7/3.7

Table 6. Ablation results of CoT filtering and transferability.

4.4. Evaluation on Rationales

To verify whether the model has effectively learned multi-step reasoning through CoTs distillation, we analyze the rationales generated by the model. Specifically, we extract and evaluate the temporal and spatial information embedded within these rationales. This approach extends beyond merely assessing the correctness of the final answer, which could be influenced by biases or other external factors. By examining the reasoning process in detail, it enables a more accurate understanding of the model's ability to perceive and reason about spatial and temporal relationships.

Evaluation protocols. We randomly select 200 samples from the STAR validation set and run inference on them using the suffix prompt, recording the generated rationales. From these rationales, we extract the predicted temporal windows and bounding boxes, comparing them to the ground truth. For the spatial part, we calculate the IoU between the predicted and ground truth bounding boxes. For the temporal part, we compute IoU and Recall, leveraging the frame-level annotations provided in the dataset.

Evaluation results. Table 7 presents the evaluation results. For comparison, we also test UniVTG for temporal reasoning and OWL-ViT v2 for spatial reasoning. The results show that LNV-Instruct struggles to generate valid rationales, even when using the suffix prompt. In contrast, LNV-AoTD demonstrates comparable performance to specialized models in both spatial and temporal reasoning, indicating that the model successfully acquired these abilities through the distillation process.

Model	Temporal Grounding IoU (%)	Temporal Grounding Recall (%)	Spatial Grounding IoU (%)
UniVTG	22.8	31.0	-
OWL-ViT v2	-	-	64.7
LNV-Instruct	✗	✗	✗
LNV-AoTD	21.7	34.0	45.2

Table 7. Temporal and spatial abilities evaluation results.

5. Conclusion

We present Agent-of-Thoughts Distillation (AoTD), that aims to distill multi-step reasoning and spatial-temporal understanding into a large video-language model (Video-LLM). Our method introduces an agent-based system that automates the generation of Chain-of-Thoughts (CoTs) from various VideoQA datasets, by breaking down complex questions into manageable sub-tasks that can be addressed by specialized vision models. Extensive experiments validate that the distilled model significantly enhances performance on both MC-VQA and OE-VQA benchmarks, underscoring the effectiveness of our approach. We believe AoTD represents a promising future direction for advancing the reasoning abilities in Video-LLMs.

Acknowledgments

This work is funded by National Key R&D Program of China (No.2022ZD0161400).

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, 2022. 1

[2] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 4

[3] Li Bo, Zhang Yuanhan, Guo Dong, Zhang Renrui, Li Feng, Zhang Hao, Zhang Kaichen, Li Yanwei, Liu Ziwei, and Li Chunyuan. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 7

[4] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024. 2, 7, 8

[5] Rohan Choudhury, Koichiro Niinuma, Kris M. Kitani, and Laszlo A. Jeni. Zero-shot video question answering with procedural programs. *arXiv preprint arXiv:2312.00937*, 2023. 2

[6] Guo Daya, Zhu Qihao, Yang Dejian, Dong Zhenda Xie, Kai, Zhang Wentao, Chen Guanting, Bi Xiao, Y. Wu, Y.K. Li, Luo Fuli, and Liang Yingfei, Xiongand Wenfeng. Deepseek-coder: When the large language model meets programming – the rise of code intelligence. *arXiv preprint arXiv:2401.14196*, 2024. 4

[7] Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. Videoagent: A memory-augmented multimodal agent for video understanding. *arXiv preprint arXiv:2403.11481*, 2024. 1

[8] 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 *Proceedings of the International Conference on Machine Learning*, 2024. 3

[9] Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024. 6

[10] Minghe Gao, Shuang Chen, Liang Pang, Yuan Yao, Jisheng Dang, Wenqiao Zhang, Juncheng Li, Siliang Tang, Yueting Zhuang, and Tat-Seng Chua. Fact: Teaching mllms with faithful, concise and transferable rationales. In *ACM Multimedia*, 2024. 2, 5, 1

[11] Timin Gao, Peixian Chen, Mengdan Zhang, Chaoyou Fu, Yunhang Shen, Yan Zhang, Shengchuan Zhang, Xiawu Zheng, Xing Sun, Liujuan Cao, et al. Cantor: Inspiring multimodal chain-of-thought of mllm. In *ACM Multimedia*, 2024. 2

[12] Gemini Team Google. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 7, 8

[13] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 6

[14] Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. Agqa: A benchmark for compositional spatio-temporal reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 6

[15] Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 1, 2

[16] 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. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2024. 2, 6, 1

[17] Ziniu Hu, Ahmet Iscen, Chen Sun, Kai-Wei Chang, Yizhou Sun, David Ross, Cordelia Schmid, and Alireza Fathi. Avis: Autonomous visual information seeking with large language model agent. In *Advances in Neural Information Processing Systems*, 2024. 1

[18] De-An Huang, Shijia Liao, Subhashree Radhakrishnan, Hongxu Yin, Pavlo Molchanov, Zhiding Yu, and Jan Kautz. Lita: Language instructed temporal-localization assistant. In *Proceedings of the European Conference on Computer Vision*, 2024. 4

[19] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2024. 2, 6, 7, 8

[20] Liunian Harold Li*, Pengchuan Zhang*, Haotian Zhang*, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 4

[21] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *Proceedings of the European Conference on Computer Vision*, 2024. 4, 7, 8

[22] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the*

Conference on Empirical Methods in Natural Language Processinng, 2024. 1, 7, 8

[23] 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. In *Proceedings of the International Conference on Computer Vision*, 2023. 1, 4

[24] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Association for Computational Linguistics*, 2024. 1, 6, 8

[25] Ahmad Mahmood, Ashmal Vayani, Muzammal Naseer, Salman Khan, and Fahad Khan. Vurf: A general-purpose reasoning and self-refinement framework for video understanding. *arXiv preprint arXiv:2403.14743*, 2024. 8

[26] Minderer Matthias, Gritsenko Alexey, Stone Austin, Neumann Maxim, Weissenborn Dirk, Dosovitskiy Alexey, Mahendran Aravindh, Arnab Anurag, Dehghani Mostafa, Shen Zhuoran, Wang Xiao, Zhai Xiaohua, Kipf Thomas, and Houlsby Neil. Simple open-vocabulary object detection with vision transformers. In *Proceedings of the European Conference on Computer Vision*, 2022. 4

[27] Juhong Min, Shyamal Buch, Arsha Nagrani, Minsu Cho, and Cordelia Schmid. Morevqa: Exploring modular reasoning models for video question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2024. 8

[28] Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. In *Advances in Neural Information Processing Systems*, 2024. 4

[29] Arindam Mitra, Luciano Del Corro, Shweti Mahajan, Andres Codas, Clarisse Simoes, Sahaj Agarwal, Xuxi Chen, Anastasia Razdaibiedina, Erik Jones, Kriti Aggarwal, et al. Orca 2: Teaching small language models how to reason. *arXiv preprint arXiv:2311.11045*, 2023. 1

[30] Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. Compositional chain-of-thought prompting for large multimodal models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2024. 2

[31] OpenAI. Gpt-3.5-turbo system card. https://platform.openai.com/docs/models/gpt-3-5-turbo, 2023. 4

[32] OpenAI. Gpt-4v(ision) system card. https://openai.com/research/gpt-4v-system-card, 2023. 7, 8

[33] 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. In *Advances in Neural Information Processing Systems*, 2023. 1, 6

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, 2021. 8

[35] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2024. 4

[36] Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. In *Advances in Neural Information Processing Systems*, 2024. 2

[37] Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. In *Proceedings of the International Conference on Computer Vision*, 2023. 1, 2, 3

[38] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 8

[39] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Chenting Wang, Guo Chen, Baoqi Pei, Rongkun Zheng, Jilan Xu, Zun Wang, et al. Internvideo2: Scaling video foundation models for multimodal video understanding. In *Proceedings of the European Conference on Computer Vision*, 2024. 1, 4

[40] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, 2022. 2

[41] Zejia Weng, Xitong Yang, Ang Li, Zuxuan Wu, and Yu-Gang Jiang. Open-vclip: Transforming clip to an open-vocabulary video model via interpolated weight optimization. In *Proceedings of the International Conference on Machine Learning*, 2023. 1, 4

[42] Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. Star: A benchmark for situated reasoning in real-world videos. In *Advances in Neural Information Processing Systems*, 2021. 1, 4, 6

[43] 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 Conference on Computer Vision and Pattern Recognition*, 2021. 1, 6, 2

[44] Jihan Yang, Shusheng Yang, Anjali Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in Space: How Multimodal Large Language Models See, Remember and Recall Spaces. *arXiv preprint arXiv:2412.14171*, 2024. 6, 7

[45] Zongxin Yang, Guikun Chen, Xiaodi Li, Wenguan Wang, and Yi Yang. Doraemongpt: Toward understanding dynamic scenes with large language models (exemplified as a video agent). In *Proceedings of the International Conference on Machine Learning*, 2024. 2

[46] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*, 2024. 2

[47] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. CLEVRER: collision events for video representation and reasoning. In *Proceedings of the International Conference on Learning Representations*, 2020. 6

[48] Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. Self-chained image-language model for video localization and question answering. In *Advances in Neural Information Processing Systems*, 2023. 4

[49] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019. 1, 6

[50] Zhou Yu, Lixiang Zheng, Zhou Zhao, Fei Wu, Jianping Fan, Kui Ren, and Jun Yu. Anetqa: A large-scale benchmark for fine-grained compositional reasoning over untrimmed videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 6, 3

[51] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2024. 6

[52] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the International Conference on Computer Vision*, 2023. 8

[53] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the Conference on Empirical Methods in Natural Language Processinng*, 2023. 1, 8

[54] 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. https://llava-vl.github.io/blog/2024-04-30-llava-next-video/, 2024. 2, 4, 6, 7, 8

[55] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023. 2

[56] Orr Zohar, Xiaohan Wang, Yonatan Bitton, Idan Szpektor, and Serena Yeung-Levy. Video-star: Self-training enables video instruction tuning with any supervision. *arXiv preprint arXiv:2407.06189*, 2024. 3