RETHINKING SCALE: HOW MULTI-AGENT COLLABORATION ENABLES SMALLER MODELS TO RIVAL GPT-4 IN VIDEO UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

The rapid development of large language models (LLMs) has brought new perspectives to the field of video understanding. However, existing methods often rely on large-scale proprietary models, such as GPT-4, to achieve competitive performance. This paper challenges the notion that scale is the primary driver of capability by introducing RIVAL, a framework demonstrating how multi-agent collaboration enables smaller open-source models (72B or fewer) to rival their large-scale counterparts. RIVAL consists of two key components: a Multi-stage React Planner (MSRP) for structured stepwise reasoning and Multi-agent Debate Refinement (MADR) for collaborative answer generation. MSRP enhances instruction-following through precise control, while MADR improves answer quality via multi-perspective debate. Using a 72B model, our framework sets a new state-of-the-art on the EgoSchema subset with 66.8% accuracy, surpassing prior GPT-4 based methods by 6.6%. Furthermore, we demonstrate that even smaller open-source models (0.6B to 32B) across the Qwen 2.5 and 3 series achieve competitive performance with RIVAL. We also demonstrate competitive performance on the Next QA benchmark. Highlighting its efficiency, RIVAL can process over 28 hours of continuous video input using limited computational resources.

1 INTRODUCTION

With the rapid advancement of multimedia technologies, video understanding has emerged as one of the key tasks in computer vision and has garnered significant attention (Gong et al., 2025; Madan et al., 2024). The rapid development of large language models (LLMs) has brought new perspectives to the field of video understanding. For example, some studies have attempted to transform video understanding tasks into text reasoning problems, leveraging the powerful semantic relationship modeling capabilities of LLMs to infer connections between video frames (Wang et al., 2024c; Ataallah et al., 2024). However, the reliance of existing LLM-based methods on proprietary models introduces two major challenges: data privacy concerns and prohibitive resource requirements.

Data privacy concerns: VideoAgent (Wang et al., 2024c), for instance, utilizes LLMs to perform observation and reflection processes to gather information and answer specific queries. However, this single-agent architecture heavily relies on the reasoning capabilities of LLMs (Figure 1 Self-Evaluation). If the generated results deviate, it can directly lead to erroneous answers. To mitigate this issue, VideoAgent relies on large-scale commercial models such as GPT-4. However, in practical applications, exposing user data to external commercial models poses significant privacy risks. **Resource**: LLoVi (Zhang et al., 2024) converts video segments into textual inputs for LLMs and achieves significant performance improvements by leveraging the strong language reasoning ability. LLoVi requires the model's text window to accommodate the full textual description of the video. When video lengths extend to tens of minutes or even hours, this requirement becomes constrained by the context window size, posing a major bottleneck for understanding long videos (Figure 1 Summary). These limitations motivate a fundamental question: Are large-scale proprietary models truly necessary to achieve competitive performance? This paper tries to answer this question. Specifically, we conduct experiments using small-scale open-source models within a resource-constrained setting, notably limiting the context window to 15,000 tokens. Operating under these constraints, we propose RIVAL, a novel framework that achieves competitive performance with a model as small as 32B parameters.

The core architecture of RIVAL comprises two modules: Multi-stage React Planner (MSRP) and Multi-agent Debate Refinement (MADR). MSRP addresses the weaker reasoning and instruction-following abilities of small-scale models. Specifically, it first decomposes original complex task into simpler sub-tasks. Then, through multi-stage prompting, it guides model's state to transition through these sub-tasks according to pre-defined rules. Notably, within sub-task, we empower the model with retrieval tool, enabling it to fetch keyframes without processing the full video content. MADR subsequently mitigates error propagation from the multi-step reasoning process. Following MSRP, an adversarial debate is initiated in which affirmative and negative agents seek evidence from opposing perspectives to challenge and progressively refine the current answer.

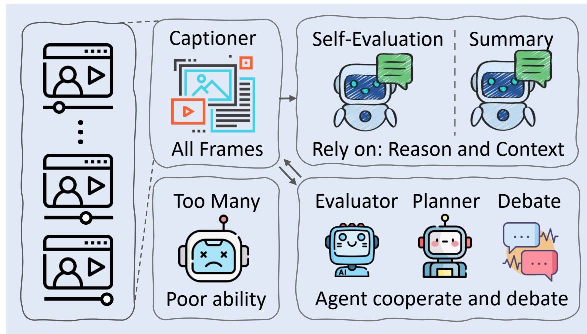

Figure 1: **Illustration of our RIVAL and prior methods.** Top: Prior methods use a single agent with large proprietary models for self-evaluation and summary, posing significant resource and privacy risks. Bottom: RIVAL leverages a cooperative multi-agent system to decompose tasks and achieve competitive performance on small models.

We evaluate RIVAL's performance against prior GPT-4-based methods on the EgoSchema (Mangalam et al., 2023) and Next QA (Xiao et al., 2021) benchmark, deploying the entire Qwen 2.5 and Qwen 3 model series. Our findings show that, despite relying on small-scale models and limited resources, RIVAL delivers competitive and often superior results. Notably, on a subset of EgoSchema, RIVAL with 72B/32B models surpasses the previous state-of-the-art by a substantial margin, achieving 66.8% and 65.0% accuracy—outperforming the GPT-4 baseline by 6.6% and 4.8%, respectively. In addition, RIVAL proves its practical scalability by handling a continuous video input exceeding 28 hours, all while operating on limited resources. The primary contributions are summarized as follows:

- We propose RIVAL, a novel video understanding framework designed to run on small-scale models in resource-constrained environments, without relying on proprietary models.
- The Multi-stage React Planner (MSRP) compensates for the weaker instruction-following of small models. MSRP decomposes complex tasks into manageable sub-tasks and enables tool use for efficient keyframe retrieval, obviating the need for full video processing.
- The Multi-agent Debate Refinement (MADR) mitigates error propagation in multi-step reasoning. This module initiates an adversarial debate between affirmative and negative agents to challenge, verify, and progressively refine answers based on evidence.
- Extensive experiments on the Qwen 2.5 and 3 series demonstrate RIVAL's superiority and practical value. The framework not only sets a new state-of-the-art on benchmarks but also proves its robustness by handling a 28-hour video, showcasing it as a high-performing, resource-efficient, and privacy-preserving solution.

2 Related Work

Long-form Video Understanding. The field of video understanding has progressed from handcrafted feature-based methods to approaches using deep learning and large language models (LLMs). Early methods relied on features like Improved Dense Trajectories (IDT) (Shu et al., 2015) and Histogram of Oriented Gradients (HOG) (Dalal & Triggs, 2005), processed by models such as Support Vector Machines (SVM) (Hearst et al., 1998). Deep learning introduced architectures like CNNs (Krizhevsky et al., 2012), LSTMs (Hochreiter & Schmidhuber, 1997), and Transformers (Vaswani et al., 2017), with notable advances including TimeSformer (Bertasius et al., 2021) and self-supervised methods like VideoMAE (Tong et al., 2022). Inspired by natural language processing, LLMs are applied to video understanding in three ways: converting videos to text for summarization (Zhao et al., 2024; Chen et al., 2023a; Xue et al., 2024); mapping video frames to text space with instruction embeddings (Shu et al., 2023; Chen et al., 2023b; Ko et al., 2023); and combining textual descriptions with video embeddings (Lin et al., 2023; Han et al., 2023; Wang et al., 2024a). Among these, the first approach is commonly used for long videos, as it alleviates visual token density issues within LLM text window constraints. However, existing methods often rely on large proprietary models, posing data privacy risks and high computational costs. To address these issues, we propose the RIVAL framework, which uses open-source, lightweight models instead of commercial LLMs.

Large Language Models Agent. Large Language Models (LLMs) are widely used for their strong contextual understanding and reasoning abilities (Meng et al., 2025; Wang et al., 2025; Liu et al., 2025). Unlike predefined, rule-driven methods (Huang et al., 2022; Dasgupta et al., 2023), agent-based systems with LLMs offer greater flexibility, enabling dynamic adaptation to complex and evolving scenarios. These agents can interact with various environments, such as games (Valmeekam et al., 2022; Yao et al., 2023b), robotics (Shridhar et al., 2021; Fan et al., 2022), and web applications (Yao et al., 2023a; Trivedi et al., 2024), efficiently integrating information for reasoning and decision-making. Agent-based approaches have demonstrated success in fields like translation (Cui et al., 2025; Chun et al., 2025), medical diagnostics (Sviridov et al., 2025; Maharana et al., 2025), and financial analysis (Lopez-Lira, 2025; Zhu et al., 2025). However, when applied to video understanding, existing methods face challenges such as dependency on extended context windows (Zhang et al., 2024). To address these issues, we propose RIVAL. RIVAL maintains stable reasoning performance and achieves performance comparable to or even surpassing proprietary large-scale commercial models.

Multi-Agent Debate. Multi-Agent Collaboration (MAC) is gradually emerging as a critical research direction in the field of natural language processing (NLP) (Chen et al., 2024; Singhal et al., 2023; Wang et al., 2024d). By facilitating information exchange, collaboration, and competition among agents, MAC effectively addresses the limitations of individual agents in reasoning capabilities, thereby significantly enhancing the overall performance and reliability of the system. Within the framework of MAC, Multi-Agent Debate (MAD) has emerged as a distinctive mode of collaboration and has garnered increasing attention (Chan et al., 2024; Smit et al., 2024; Liang et al., 2024). This approach emulates the human debating process, enabling each agent to dynamically present arguments, refute opposing viewpoints, and iteratively refine their own positions. MAD not only broadens the depth and scope of reasoning but also improves the diversity and logical consistency of generated content. Answer generation in ours RIVAL is modeled as a MAD process, where agents are empowered to autonomously retrieve relevant information to substantiate their respective arguments, thereby providing stronger evidential support. Through the debate process, the agents contribute perspectives and analyses, enhancing the comprehensiveness and objectivity of the generated responses.

3 METHOD

Our pipeline is illustrated in Figure 2. Consistent with prior work, we decode the video into frames using a fixed sampling rate. Retrieval tools are then used to generate initial information. This information is fed into an iterative process to dynamically add missing details and filter out irrelevant content. During this process, MSRP analyzes the evaluation results and generates usage plans for the retrieval tools. Finally, the updated information is used to produce an initial answer. MADR initiates a debate based on the initial answer, offering multi-perspective insights to refine the initial response.

3.1 INFORMATION INITIALIZATION

Since the video content remains almost consistent within short time intervals, we follow prior work (Zhang et al., 2024) and sample the video at a fixed frame rate to obtain a sequence of frames. VideoAgent randomly samples frames at fixed intervals. However, this approach may lead to the selection of irrelevant frames, which could impact subsequent reasoning. To address this, we employ CLIP for image-text alignment to sample frames that are most relevant to the given query:

$$\mathrm{I_s} = \mathrm{I_q} \cup \mathrm{I_a},\ \mathrm{I_q} = \mathrm{Top_k}(\mathrm{Sim}(\mathrm{I}, \mathrm{Q})),\ \mathrm{I_a} = \mathrm{Top_k}(\mathrm{Sim}(\mathrm{I}, \mathrm{A})). \tag{1}$$

Here, I represents the images decoded from the video. Q and A denote the question and the option, respectively. Sim denotes the image-text similarity score, and $\mathrm{Top_k}$ refers to selecting the top k frames with the highest similarity. To extract image information, we use an image description model to generate textual descriptions. This retrieve-then-describe approach reduces the need to process all frames while ensuring an effective starting point for the information retrieval process.

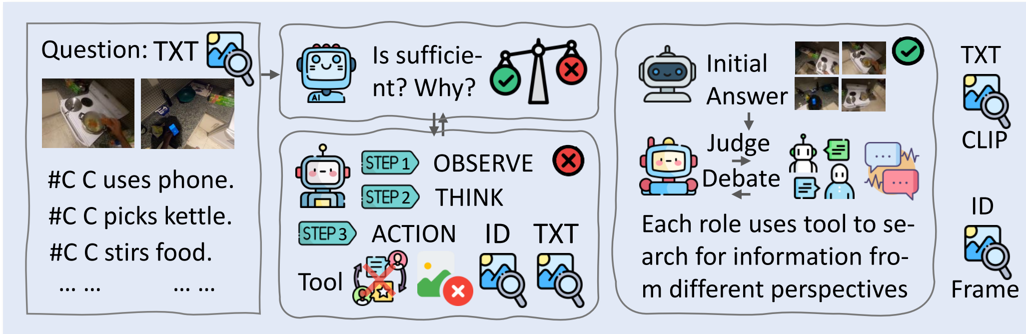

Figure 2: **The pipeline of our RIVAL**. The raw video is decoded into frames, and retrieval tools are used to extract initial information (left). MSRP interacts with evaluators to dynamically add missing information and filter out irrelevant content (middle). Finally, MADR generates an initial answer based on the collected information and initiates a debate process to refine the answer further (right).

3.2 Multi-stage React Planner

In the information initialization phase, we retrieved some information relevant to the problem. However, two main issues remain: (1) the information generated lacks clear distinctions between the characteristics and trade-offs of candidate options, increasing uncertainty in decision-making, and (2) it fails to effectively filter out noise (e.g., irrelevant or low-quality content). To address these issues, we designed a dual-agent collaborative mechanism consisting of an Evaluator and a Planner (Figure 2 middle). This mechanism aims to extract key information and improve decision-making accuracy.

Evaluator. The Evaluator is responsible for assessing the collected information on a scale of 1 to 10, where 10 represents the highest quality. It evaluates two aspects—clarity and alignment with problem-related information—weighted at 60% and 40%, respectively. Furthermore, the Evaluator provides explanatory feedback alongside its scores to better guide the Planner in refining the process:

$$\mathrm{Evaluation}(\mathrm{I_s}) = \mathrm{LLM}(\mathrm{Prompt_e}(\mathrm{M_c}(\mathrm{I_s}), \mathrm{Accuracy}, \mathrm{Completeness})), \tag{2}$$

where $\mathrm{M_c}$ denotes the description model, which is a large model that generates textual descriptions corresponding to input images. $\mathrm{Prompt_e}$ indicates the pre-defined evaluation instructions tailored to guide the model's output. $\mathrm{Accuracy}$ and $\mathrm{Completeness}$ represent the evaluation criteria, with weights assigned at 60% and 40%, respectively, reflecting their relative importance in the assessment process. The result $\mathrm{Evaluation}(\mathrm{I_s})$ consists of two elements: score, which quantifies the evaluation results numerically, and reason, which provides a justification or explanation for the assigned score.

Planner (MSRP). The Planner's main goal is to search for relevant information and remove noise data, providing clearer and more complete input for the task. In our framework, the Planner works with the Evaluator to create plans that maximize the Evaluator's score, ensuring alignment with task requirements. However, current LLMs show limitations when using the ReAct (Yao et al., 2023c), unlike commercial-grade models. For example, they may skip deep reasoning when making plans or stop at reasoning without taking actions, which reduces task efficiency. To address these challenges, we propose a Multi-Stage ReAct Planner (MSRP). MSRP decomposes the reasoning and action phases of the ReAct into multiple predefined sub-stages, forcing the model to progress step-by-step through the stages while performing explicit state transitions. This approach effectively mitigates inconsistencies between the reasoning and action processes and enhances the quality of plan:

$$\mathrm{State_{THINK}} = \mathrm{LLM}(\mathrm{Prompt_p}(\mathrm{State_{init}}, \mathrm{Evaluation}(\mathrm{I_s}), \mathrm{OBSERVE})). \tag{3}$$

In the $\mathrm{Prompt_p}$, we instruct the LLM to generate plans in three distinct phases: OBSERVE, THINK, and ACT. The initial state, denoted as $\mathrm{State_{init}}$, represents the starting condition for generating the OBSERVE phase. Since OBSERVE is the first phase, $\mathrm{State_{init}}$ is initialized as empty. OBSERVE specifies the task to be performed in the current state. We can replace $\mathrm{State_{init}}$ with $\mathrm{State_{THINK}}$ and assign the task of THINK to generate $\mathrm{State_{ACT}}$. In the final state, we provide the LLM with a list of callable tools and instruct it to generate a usage plan based on the available tools:

$$\mathrm{Plan} = \mathrm{LLM}(\mathrm{Prompt_p}(\mathrm{State_{ACT}}, \mathrm{Evaluation}(\mathrm{I_s}), \mathrm{ACT}, \mathrm{Tool_p})), \tag{4}$$

where we define four tools ($\mathrm{Tool_p}$): (1) Stop Searching, which terminates the search loop prematurely instead of waiting until the maximum number of searches is reached; (2) Delete by Frame ID, which removes irrelevant information to enhance the clarity of the collected data; (3) Add by Frame ID, which extracts images corresponding to the specified frame ID, generates descriptions using $\mathrm{M_c}$, and adds them; and (4) Add by Text, which queries CLIP to match the most similar frame based on the input text, generates descriptions using $\mathrm{M_c}$, and adds them. By generating a structured tool-usage plan, MSRP can refine the task-related information, enhancing the accuracy of subsequent answers. Subsequently, the plan is executed through function calls to modify the content of the information.

This process of evaluation, planning, and execution is repeated iteratively until a predefined condition is satisfied. Specifically, we define three conditions for termination: (1) the Planner calls the Stop Searching method for the loop, (2) the score assigned by the Evaluator exceeds a predefined threshold, or (3) the maximum number of iteration steps is reached. The loop ends when any condition is met:

$$\mathrm{Cond_{lp}} = (\mathrm{Score_e} > \alpha) \vee (\mathrm{Call\ Stop\ Searching}) \vee (\mathrm{Loop\ Number} > \mathrm{Max\ Number}). \quad (5)$$

3.3 Multi-agent Debate Refinement

The Evaluator collaborates with the Planner to generate clear and precise descriptions of the problem. We utilize this information to address the given question. As shown in Figure 2 (right), we first employ an agent to generate an initial answer. However, compared to the commercial model, the initial answer exhibits significant shortcomings in logic and consistency. Furthermore, while our approach of using step-by-step state transitions for reasoning is highly efficient, it increases the risk of error accumulation. To address these issues, we propose the Multi-Agent Debate Refinement (MADR) to refine and optimize the initial answer. Through MADR, Agents with different roles analyze the initial answer from multiple perspectives, incorporating diverse viewpoints to supplement and correct it. This multi-perspective refinement process directly corrects logical fallacies, resolves inconsistencies, and mitigates the accumulation of errors from the initial reasoning stages.

Initial Answer. In a straightforward manner, we consolidate all previously collected information into the Agent's prompt, enabling it to generate an answer. Similar to the Evaluator, we require the Agent to provide the reason behind its answer when generating the response:

$$\mathrm{Initial\ Answer} = \mathrm{LLM}(\mathrm{Prompt_{IA}}(\mathrm{M_c}(\mathrm{I_{IA}}))), \quad (6)$$

where $\mathrm{Prompt_{IA}}$ represents the instruction provided to the LLM to given the answer. $\mathrm{I_{IA}}$ denotes the frame IDs obtained during the prior information retrieval phase.

Responder (MADR). Given an initial answer, we initiate a debate process to refine it. Our debate process consists of three core roles: the affirmative side, the opposition, and the judge. Within this framework, the affirmative side is responsible for supporting the given answer, while the opposition is tasked with challenging it. When presenting counterarguments, the opposition is required to propose an alternative option. If the opposition prevails, the proposed alternative replaces the initial answer:

$$\mathrm{Statement} = \mathrm{LLM}(\mathrm{Prompt_{db}}(\mathrm{M_c}(\mathrm{I_{IA}}), \mathrm{Task}, \mathrm{Statement_{op}}, \mathrm{Tool_{db}}), \quad (7)$$

where Task refers to the role-specific objectives. The affirmative side is responsible for supporting the given answer, while the opposition is tasked with opposing it. $\mathrm{Statement_{op}}$ represents the statement generated by the opposing role. In $\mathrm{Prompt_{db}}$, we require both the affirmative side and opposition to consider the opposing arguments when generating their statements. Additionally, they are prompted to utilize the tool $\mathrm{Tool_{db}}$ to search for evidence supporting their respective positions.

We allow both the affirmative side and the opposition to call the tool only once per response. The list of available tools contains two options: (1) Frame ID Query Tool. This tool retrieves the description corresponding to a given frame ID by extracting the image associated with the ID and then generating a description using $\mathrm{M_c}$. (2) Text Query Tool. This tool uses CLIP to retrieve the frame ID corresponding to a given textual query and then generates a description using $\mathrm{M_c}$. Finally, the judge evaluates the statements presented by both the affirmative side and the opposition during each round of the debate. The debate terminates when the judge determines that either a consensus has been reached or one side has prevailed. Otherwise, the debate continues until the pre-defined maximum number of rounds is reached. Thus, the conditions for terminating the debate are as follows:

$$\mathrm{Cond_{db}} = \mathrm{Agreement} \vee \mathrm{Win} \vee (\mathrm{Number\ Round} > \mathrm{Max\ Number}), \quad (8)$$

Table 1: **Comparison with other state-of-the-art methods on the EgoSchema benchmark.** We provide two sets of comparisons. The left side presents the performance of our RIVAL method compared to other approaches based on LLM/vLLM, while the right side outlines comparisons with other training-based or large-scale proprietary models. #parm indicates the number of parameters.

Method	Model	Subset	Full
MoReVQA (Min et al., 2024)	PaLM 2	-	51.7
ProViQ (Choudhury et al., 2023)	GPT 3.5	57.1	-
IG-VLM (Kim et al., 2024)	GPT 4V	59.8	-
MVU (Ranasinghe et al., 2024)	LLaVA	13 B	-
ViViT (Papalampidi et al., 2024)	ViViT	56.8	33.3
SeViLA (Yu et al., 2023)	BLIP-2	25.7	22.7
Vamos (Wang et al., 2024b)	GPT 4	51.2	48.3
LLoVi (Zhang et al., 2024)	GPT 4	57.6	50.3
VideoAgent (Wang et al., 2024c)	GPT 4	60.2	54.1
Ours: RIVAL (72B)	Qwen 2.5	**66.8**	56.4
Ours: RIVAL (32B)	Qwen 3	65.0	**57.2**

Method	Subset	Full
Random Chance	20.0	20.0
Bard + ImageViT	35.0	35.0
+ ShortViViT	42.0	36.2
+ PALI (Papalampidi et al., 2024)	44.8	39.2
FrozenBiLM (Yang et al., 2022)	-	26.9
InternVideo (Wang et al., 2022)	-	32.1
GPT 4 Turbo	31.0	30.8
GPT 4V (Wang et al., 2024c)	63.5	55.6
Gemini 1.0 Pro (Team et al., 2024)	-	55.7
Ours: RIVAL (Qwen 2.5)	**66.8**	56.4
Ours: RIVAL (Qwen 3)	65.0	**57.2**

where $\mathrm{Agreement}$ and Win represent two judgment outcomes made by the judge: $\mathrm{Agreement}$ indicates that both sides have reached a consensus, while Win signifies that one side has prevailed over the other. $\mathrm{Number\ Round}$ refers to the number of rounds in the debate. If the debate ends without reaching a consensus, meaning the maximum number of rounds has been reached, the judge must choose a winner from either the affirmative side or the opposition.

The pseudocode for MSRP and MADR is provided in Appendix A (Algorithm 1 and 2).

4 Experiments

4.1 Details

Consistent with prior work Zhang et al. (2024); Wang et al. (2024c), we sample video frames at 1-second intervals. For experiments conducted on EgoSchema, we utilize LaViLa Zhao et al. (2023) as the video description model. To prevent data leakage, we employ a version of the model trained on Ego4D with all segments overlapping with EgoSchema removed. For Next-QA, we leverage CogAgent as the description model, aligning with previous studies Zhang et al. (2024); Wang et al. (2024c). As the retrieval model, we adopt EVA-CLIP-8B plus Sun et al. (2023), using the frame with the highest cosine similarity as the matched item during the retrieval process. For the LLM, we utilize the open-source Qwen-2.5/-3 series for all experiments. To ensure compatibility with OpenAI's standard workflow, we utilize vllm Kwon et al. (2023) to deploy the LLM on an A100 GPU (80GB) and configure the maximum number of tokens for inference to 15,000. Furthermore, due to the inability of the 72B model to be deployed on a single A100 GPU, we leverage vllm to enable tensor parallelism and perform inference across two A100 GPUs (totaling 160GB). All weights of Qwen2.5 are obtained from the official repository of the ModelScope community. For all model scales, we utilize instruction-tuned versions, such as Qwen2.5-72B-instruct.

4.2 Main Result

EgoSchema. In Table 1, we present the performance comparison between our RIVAL method and other SOTA method on the EgoSchema benchmark. Compared to smaller-scale methods such as MoReVQA, ProViQ, and IG-VLM, our RIVAL demonstrates significant performance advantages. On the subset, RIVAL achieves an accuracy improvement of 10 points over ViViT and 31.1 points over SeViLA. Similarly, RIVAL maintains its superiority with an 11-point accuracy margin ahead of MVU. When compared to approaches based on GPT 4, our method also showcases competitive advantages. On the subset, RIVAL outperforms Vamos by 15.6 points, LLoVi by 9.2 points, and VideoAgent by 6.6 points. On the full dataset, RIVAL achieves consistent superiority with an accuracy margin of 8.1, 6.1, and 2.3 points over Vamos, LLoVi, and VideoAgent, respectively. Even when compared to proprietary models, RIVAL holds a leading position. On the subset and full dataset, RIVAL surpasses

Table 2: **Comparison with other state-of-the-art (SOTA) methods on the Next QA benchmark.** The ATP-Hard subset represents the validation set, which is a more challenging subset. The upper portion of the table presents methods based on supervised training, while the lower portion lists methods utilizing LLM/vLLM. Our RIVAL demonstrates significant performance advantages.

Method	#parm	Validation Set				ATP-hard subset		
		All	Causual	Temporal	Descriptive	All	Causual	Temporal
Random Chance	-	20	20	20	20	20	20	20
VFC (Yang et al., 2021)	164 M	52.3	49.6	51.5	63.2	-	-	-
ATP (Buch et al., 2022)	88 M	54.3	53.1	50.2	66.8	38.8	38.4	36.5
MIST (Gao et al., 2023b)	88 M	57.2	54.6	56.6	66.9	-	-	-
GF (Bai et al., 2024)	88 M	58.8	56.9	57.1	70.5	49.3	48.7	50.3
CoVGT (Xiao et al., 2023)	149 M	60.7	59.7	58.0	69.9	-	-	-
SeViT (Kim et al., 2023)	215 M	56.7	54.0	54.1	71.3	-	43.3	46.5
HiTeA (Ye et al., 2023)	297 M	63.1	62.4	58.3	75.6	-	47.8	48.6
VFC (Yang et al., 2021)	540 B	51.5	51.6	45.4	64.1	31.4	32.2	30.0
InternVideo (Wang et al., 2022)	-	49.1	43.4	48.0	65.1	-	-	-
AssistGPT (Gao et al., 2023a)	1.8 T	58.4	60.0	51.4	67.3	-	-	-
ViperGPT (Surís et al., 2023)	175B	60.0				-	-	-
SeViLA (Yu et al., 2023)	4 B	63.6	61.3	61.5	75.6	-	-	-
LLoVi (Zhang et al., 2024)	1.8 T	67.7	69.5	61.0	75.6	-	-	-
VideoAgent (Wang et al., 2024c)	1.8 T	71.3	72.7	64.5	81.1	58.4	57.8	58.8
Ours:RIVAL (Qwen 2.5)	72 B	**74.4**	**76.3**	**67.5**	82.4	**66.5**	**70.2**	**61.2**
Ours:RIVAL (Qwen 3)	32 B	73.2	74.1	67.0	**82.9**	63.7	66.2	60.1

GPT 4V by 3.3 and 0.8 points, respectively. Furthermore, RIVAL is training-free, yet it achieves significant performance gains over other training-based approaches. Specifically, RIVAL outperforms FrozenBiLM (Yang et al., 2022) by 25.5 points and InterVideo (Wang et al., 2022) by 23.3 points.

Next-QA. Table 2 presents the performance comparison between our RIVAL method and other SOTA approaches on the Next-QA benchmark. Compared to supervised training-based methods, RIVAL demonstrates significant advantages. Specifically, when compared to methods using the ViT-B-32 (Dosovitskiy et al., 2021) backbone, including ATP (Buch et al., 2022), MIST (Gao et al., 2023b), and GF (Bai et al., 2024), our RIVAL achieves performance improvements of 20.1, 17.3, and 15.6 points respectively. On the more challenging ATP-Hard subset, the performance gap further widens, with RIVAL outperforming ATP and GF by 27.7 and 17.2 points, respectively, highlighting its superior capability in difficult scenarios. When compared to methods based on LLM/vLLM, RIVAL also achieves considerable performance margins, outperforming VFC and SeViLA (Yu et al., 2023) by 22.9 and 10.8 points, respectively. Even in comparison with approaches utilizing GPT 4, our RIVAL maintains its leading position. On the validation set, RIVAL achieves accuracy improvements of 18.3 points over AssistGPT (Gao et al., 2023a), 7 points over LLoVi, and 3.1 points over VideoAgent. Furthermore, on the challenging subset, RIVAL achieves even more substantial performance differences, surpassing VFC and VideoAgent by 35.1 and 8.1 points, respectively.

4.3 Comparison Across Various Scales

To further evaluate the practical performance of our RIVAL, we present in Table 3 a performance comparison across various scales. Notably, to assess performance on long videos, we concatenated all videos from the EgoSchema subset to form a single video approximately 28 hours in length. Based on this long video, we evaluated the performance of answering questions from the EgoSchema subset, which is labeled as Long in Table 3. Furthermore, we implemented VideoAgent using the full Qwen 2.5 model series (see Appendix C). First, RIVAL demonstrates remarkable capital efficiency. Across both the Qwen 2.5 and 3 series, it consistently matches the performance of prior methods like LLoVi and VideoAgent but with significantly smaller models. For instance, on the Qwen 3 series, a 1.7B RIVAL model can achieve performance comparable to much larger counterparts. Second, in direct comparisons on identical models, RIVAL consistently and significantly outperforms VideoAgent. This performance gap widens on more challenging tasks. On the EgoSchema evaluation dataset, for

Table 3: **Performance comparison across various scales of LLM.** Scale indicates the parameter size of the LLM, all of which are from the Qwen 2.5/3 series. Subset refers to the validation set of EgoSchema. All videos in EgoSchema were concatenated into a single approximately **28-hour-long** video to evaluate the performance on all questions in the Subset, Long (28h). Additionally, we provide implementation of VideoAgent using the Qwen 2.5 series in Appendix C.

Scale	EgoSchema		Next QA (Val)				Next QA (ATP-Hard)		
	Long (28h)	Subset	All	Causual	Temporal	Descriptive	All	Causual	Temporal
Qwen 2.5 Series: 14 B ≈ LLoVi; 32 B ≈ VideoAgent									
72 B	48.7	66.8	74.4	76.3	67.5	82.4	66.5	70.2	61.2
32 B	44.8	61.4	72.4	74.4	65.1	80.4	63.9	68.0	58.0
14 B	45.2	57.6	70.5	72.4	63.3	78.4	62.2	65.9	56.8
7 B	41.4	53.2	66.9	68.2	59.9	76.2	58.1	61.4	53.3
3 B	38.9	53.0	59.0	59.2	54.8	67.1	50.3	53.0	46.6
1.5 B	33.8	43.6	53.6	55.2	49.8	56.6	45.9	49.3	41.1
Qwen 3 Series: 1.7 B ≈ LLoVi; 8 B ≈ VideoAgent									
32 B	46.0	65.0	73.2	74.1	67.0	82.9	63.7	66.2	60.1
14 B	46.2	60.4	71.9	72.7	66.3	80.7	62.7	65.3	58.9
8 B	47.1	60.4	70.8	72.4	63.9	79.5	61.5	64.3	57.5
4 B	45.2	59.2	68.8	69.5	63.0	78.3	59.7	61.9	56.7
1.7 B	41.3	56.7	62.9	63.0	57.7	57.7	52.7	54.5	50.1
0.6 B	35.7	45.6	53.9	53.3	49.8	64.5	45.1	47.3	41.8

example, RIVAL achieves up to a 10-point accuracy improvement over VideoAgent (using Qwen 2.5 72B). Finally, RIVAL shows superior robustness in extreme-length video scenarios. On the 28-hour concatenated video, VideoAgent's performance collapses to near-random levels. In stark contrast, RIVAL maintains robust performance with only minor degradation, outperforming VideoAgent by as much as 14.8 points and proving its capability for real-world, long-form video analysis.

4.4 Case Study

In Figure 3, we present a case study comparing RIVAL and VideoAgent. VideoAgent follows a response-reflection process. However, when the initial response contains errors, the subsequent reflection process is prone to failure. As shown in the middle section of Figure 3, the initial response provides an incorrect answer. During the subsequent reflection step, LLM assigns a confidence score of 3 to the answer, indicating that it fails to recognize the mistake in its generated response. Consequently, the final answer reproduces the same error. In contrast, our RIVAL framework successfully identifies and corrects this error. After the initial information retrieval phase, MSRP conducts further searches to gather additional information relevant to the given question. However, similar to VideoAgent, RIVAL also generates an incorrect initial response. This error is later identified and resolved during the subsequent MADR process. Specifically, the counterarguments generated during the debate process successfully identify the inconsistency. The counterarguments highlight that, based on the observed information, no details link entity C's behavior to Option 1, and they provide a more plausible alternative answer. The judge module, considering the counterarguments, determines that the counterarguments are more persuasive, revising the final answer accordingly.

5 Conclusion

This paper proposed RIVAL, a novel framework that demonstrates how smaller, open-source models can achieve competitive and often state-of-the-art performance in video understanding by addressing the key limitations of current methods. Through its two components—MSRP for structured, tool-assisted reasoning and MADR for adversarial answer refinement—RIVAL operates effectively as a training-free solution. Crucially, RIVAL circumvents the primary bottlenecks for practical application: its ability to process video without full-frame ingestion mitigates context window limitations, while its independence from proprietary APIs addresses significant data privacy concerns. This allows

Figure 3: **Case Study of VideoAgent and RIVAL on EgoSchema Benchmark (Qwen 2.5 72B).** The top part displays the given question and corresponding options. In the middle section, the reasoning process of VideoAgent is presented, which highlights its step-by-step analysis approach. Despite including a self-assessment process, it fails to identify its own errors. The bottom section illustrates our RIVAL framework. RIVAL first retrieves information and then initiates a debate process based on the retrieved information, generating opinions from diverse perspectives to refine the answer.

RIVAL to robustly handle videos of arbitrary and even extreme length, paving the way for more accessible, efficient, secure, and truly scalable video understanding in real-world scenarios.

6 Reproducibility Statement

Due to our company's data security policies, we are currently unable to release the full source code. Upon completion of the internal security review, we plan to make the corresponding experimental logs and replication scripts publicly available. In the interim, to facilitate reproducibility, we provide detailed pseudocode for both MSRP and MADR in the Appendix. Furthermore, all prompt templates used in our study are detailed in the subsequent sections. For dataset configurations and content descriptions, we refer readers to the official open-source repositories of VideoAgent and LLoVi. We believe these resources provide a sufficient basis for replicating our key findings.

References

Kirolos Ataallah, Xiaoqian Shen, Eslam Abdelrahman, Essam Sleiman, Mingchen Zhuge, Jian Ding, Deyao Zhu, Jürgen Schmidhuber, and Mohamed Elhoseiny. Goldfish: Vision-language understanding of arbitrarily long videos, 2024. URL https://arxiv.org/abs/2407.12679.

Ziyi Bai, Ruiping Wang, and Xilin Chen. Glance and focus: Memory prompting for multi-event video question answering. *CoRR*, abs/2401.01529, 2024. doi: 10.48550/ARXIV.2401.01529. URL https://doi.org/10.48550/arXiv.2401.01529.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 813–824. PMLR, 2021. URL http://proceedings.mlr.press/v139/bertasius21a.html.

Shyamal Buch, Cristobal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, and Juan Carlos Niebles. Revisiting the "Video" in Video-Language Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. Chateval: Towards better llm-based evaluators through multi-agent debate. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=FQepisCUWu.

Justin Chih-Yao Chen, Swarnadeep Saha, Elias Stengel-Eskin, and Mohit Bansal. Magdi: Structured distillation of multi-agent interaction graphs improves reasoning in smaller language models. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=ffLblkoCw8.

Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Mingzhen Sun, Xinxin Zhu, and Jing Liu. VAST: A vision-audio-subtitle-text omni-modality foundation model and dataset. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023a. URL http://papers.nips.cc/paper_files/paper/2023/hash/e6b2b48b5ed90d07c305932729927781-Abstract-Conference.html.

Tao Chen, Enwei Zhang, Yuting Gao, Ke Li, Xing Sun, Yan Zhang, and Hui Li. MMICT: boosting multi-modal fine-tuning with in-context examples. *CoRR*, abs/2312.06363, 2023b. doi: 10.48550/ARXIV.2312.06363. URL https://doi.org/10.48550/arXiv.2312.06363.

Rohan Choudhury, Koichiro Niinuma, Kris M. Kitani, and László A. Jeni. Zero-shot video question answering with procedural programs. *CoRR*, abs/2312.00937, 2023. doi: 10.48550/ARXIV.2312.00937. URL https://doi.org/10.48550/arXiv.2312.00937.

Jina Chun, Qihong Chen, Jiawei Li, and Iftekhar Ahmed. Is multi-agent debate (mad) the silver bullet? an empirical analysis of mad in code summarization and translation, 2025. URL https://arxiv.org/abs/2503.12029.

Bowen Cui, Tejas Ramesh, Oscar Hernandez, and Keren Zhou. Do large language models understand performance optimization?, 2025. URL https://arxiv.org/abs/2503.13772.

N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pp. 886–893 vol. 1, 2005. doi: 10.1109/CVPR.2005.177.

Ishita Dasgupta, Christine Kaeser-Chen, Kenneth Marino, Arun Ahuja, Sheila Babayan, Felix Hill, and Rob Fergus. Collaborating with language models for embodied reasoning. *CoRR*, abs/2302.00763, 2023. doi: 10.48550/ARXIV.2302.00763. URL https://doi.org/10.48550/arXiv.2302.00763.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Linxi Fan, Guanzhi Wang, Yunfan Jiang, Ajay Mandlekar, Yuncong Yang, Haoyi Zhu, Andrew Tang, De-An Huang, Yuke Zhu, and Anima Anandkumar. Minedojo: Building open-ended embodied agents with internet-scale knowledge. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/74a67268c5cc5910f64938cac4526a90-Abstract-Datasets_and_Benchmarks.html.

Difei Gao, Lei Ji, Luowei Zhou, Kevin Qinghong Lin, Joya Chen, Zihan Fan, and Mike Zheng Shou. Assistgpt: A general multi-modal assistant that can plan, execute, inspect, and learn. *CoRR*, abs/2306.08640, 2023a. doi: 10.48550/ARXIV.2306.08640. URL https://doi.org/10.48550/arXiv.2306.08640.

Difei Gao, Luowei Zhou, Lei Ji, Linchao Zhu, Yi Yang, and Mike Zheng Shou. MIST : Multi-modal iterative spatial-temporal transformer for long-form video question answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 14773–14783. IEEE, 2023b. doi: 10.1109/CVPR52729.2023.01419. URL https://doi.org/10.1109/CVPR52729.2023.01419.

Linxiao Gong, Hao Yang, Gaoyun Fang, Bobo Ju, Juncen Guo, Xiaoguang Zhu, Xiping Hu, Yan Wang, Peng Sun, and Azzedine Boukerche. A survey on video analytics in cloud-edge-terminal collaborative systems, 2025. URL https://arxiv.org/abs/2502.06581.

Mingfei Han, Linjie Yang, Xiaojun Chang, and Heng Wang. Shot2story20k: A new benchmark for comprehensive understanding of multi-shot videos. *CoRR*, abs/2312.10300, 2023. doi: 10.48550/ARXIV.2312.10300. URL https://doi.org/10.48550/arXiv.2312.10300.

M.A. Hearst, S.T. Dumais, E. Osuna, J. Platt, and B. Scholkopf. Support vector machines. *IEEE Intelligent Systems and their Applications*, 13(4):18–28, 1998. doi: 10.1109/5254.708428.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8): 1735–1780, November 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.8.1735. URL https://doi.org/10.1162/neco.1997.9.8.1735.

Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. *CoRR*, abs/2201.07207, 2022. URL https://arxiv.org/abs/2201.07207.

Sungdong Kim, Jin-Hwa Kim, Jiyoung Lee, and Minjoon Seo. Semi-parametric video-grounded text generation. *CoRR*, abs/2301.11507, 2023. doi: 10.48550/ARXIV.2301.11507. URL https://doi.org/10.48550/arXiv.2301.11507.

Wonkyun Kim, Changin Choi, Wonseok Lee, and Wonjong Rhee. An image grid can be worth a video: Zero-shot video question answering using a VLM. *CoRR*, abs/2403.18406, 2024. doi: 10.48550/ARXIV.2403.18406. URL https://doi.org/10.48550/arXiv.2403.18406.

Dohwan Ko, Ji Soo Lee, Woo-Young Kang, Byungseok Roh, and Hyunwoo Kim. Large language models are temporal and causal reasoners for video question answering. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pp. 4300–4316. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.EMNLP-MAIN.261. URL https://doi.org/10.18653/v1/2023.emnlp-main.261.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. Encouraging divergent thinking in large language models through multi-agent debate. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pp. 17889–17904. Association for Computational Linguistics, 2024. URL https://aclanthology.org/2024.emnlp-main.992.

Kevin Lin, Faisal Ahmed, Linjie Li, Chung-Ching Lin, Ehsan Azarnasab, Zhengyuan Yang, Jianfeng Wang, Lin Liang, Zicheng Liu, Yumao Lu, Ce Liu, and Lijuan Wang. MM-VID: advancing video understanding with gpt-4v(ision). *CoRR*, abs/2310.19773, 2023. doi: 10.48550/ARXIV.2310.19773. URL https://doi.org/10.48550/arXiv.2310.19773.

Hui Liu, Wenya Wang, Kecheng Chen, Jie Liu, Yibing Liu, Tiexin Qin, Peisong He, Xinghao Jiang, and Haoliang Li. Enhancing zero-shot image recognition in vision-language models through human-like concept guidance, 2025. URL https://arxiv.org/abs/2503.15886.

Alejandro Lopez-Lira. Can large language models trade? testing financial theories with llm agents in market simulations, 2025. URL https://arxiv.org/abs/2504.10789.

Neelu Madan, Andreas Moegelmose, Rajat Modi, Yogesh S. Rawat, and Thomas B. Moeslund. Foundation models for video understanding: A survey, 2024. URL https://arxiv.org/abs/2405.03770.

Umakanta Maharana, Sarthak Verma, Avarna Agarwal, Prakashini Mruthyunjaya, Dwarikanath Mahapatra, Sakir Ahmed, and Murari Mandal. Right prediction, wrong reasoning: Uncovering llm misalignment in ra disease diagnosis, 2025. URL https://arxiv.org/abs/2504.06581.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding, 2023. URL https://arxiv.org/abs/2308.09126.

Yuan Meng, Xiangtong Yao, Haihui Ye, Yirui Zhou, Shengqiang Zhang, Zhenshan Bing, and Alois Knoll. Data-agnostic robotic long-horizon manipulation with vision-language-guided closed-loop feedback, 2025. URL https://arxiv.org/abs/2503.21969.

Juhong Min, Shyamal Buch, Arsha Nagrani, Minsu Cho, and Cordelia Schmid. Morevqa: Exploring modular reasoning models for video question answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 13235–13245. IEEE, 2024. doi: 10.1109/CVPR52733.2024.01257. URL https://doi.org/10.1109/CVPR52733.2024.01257.

Pinelopi Papalampidi, Skanda Koppula, Shreya Pathak, Justin Chiu, Joe Heyward, Viorica Patraucean, Jiajun Shen, Antoine Miech, Andrew Zisserman, and Aida Nematzadeh. A simple recipe for contrastively pre-training video-first encoders beyond 16 frames. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 14386–14397. IEEE, 2024. doi: 10.1109/CVPR52733.2024.01364. URL https://doi.org/10.1109/CVPR52733.2024.01364.

Kanchana Ranasinghe, Xiang Li, Kumara Kahatapitiya, and Michael S. Ryoo. Understanding long videos in one multimodal language model pass. *CoRR*, abs/2403.16998, 2024. doi: 10.48550/ARXIV.2403.16998. URL https://doi.org/10.48550/arXiv.2403.16998.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew J. Hausknecht. Alfworld: Aligning text and embodied environments for interactive learning. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=0IOX0YcCdTn.

Fangxun Shu, Lei Zhang, Hao Jiang, and Cihang Xie. Audio-visual LLM for video understanding. *CoRR*, abs/2312.06720, 2023. doi: 10.48550/ARXIV.2312.06720. URL https://doi.org/10.48550/arXiv.2312.06720.

Zhixin Shu, Kiwon Yun, and Dimitris Samaras. Action detection with improved dense trajectories and sliding window. In Lourdes Agapito, Michael M. Bronstein, and Carsten Rother (eds.), *Computer Vision - ECCV 2014 Workshops*, pp. 541–551, Cham, 2015. Springer International Publishing. ISBN 978-3-319-16178-5.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Andrew Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Agüera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle K. Barral, Dale R. Webster, Gregory S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. Towards expert-level medical question answering with large language models. *CoRR*, abs/2305.09617, 2023. doi: 10.48550/ARXIV.2305.09617. URL https://doi.org/10.48550/arXiv.2305.09617.

Andries P. Smit, Nathan Grinsztajn, Paul Duckworth, Thomas D. Barrett, and Arnu Pretorius. Should we be going mad? A look at multi-agent debate strategies for llms. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=CrUmgUaAQp.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.

Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 11854–11864. IEEE, 2023. doi: 10.1109/ICCV51070.2023.01092. URL https://doi.org/10.1109/ICCV51070.2023.01092.

Ivan Sviridov, Amina Miftakhova, Artemiy Tereshchenko, Galina Zubkova, Pavel Blinov, and Andrey Savchenko. 3mdbench: Medical multimodal multi-agent dialogue benchmark, 2025. URL https://arxiv.org/abs/2504.13861.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Jack Krawczyk, Cosmo Du, Ed Chi, Heng-Tze Cheng, Eric Ni, Purvi Shah, Patrick Kane, Betty Chan, Manaal Faruqui, Aliaksei Severyn, Hanzhao Lin, YaGuang Li, Yong Cheng, Abe Ittycheriah, Mahdis Mahdieh, Mia Chen, Pei Sun, Dustin Tran, Sumit Bagri, Balaji Lakshminarayanan, Jeremiah Liu, Andras Orban, Fabian Güra, Hao Zhou, Xinying Song, Aurelien Boffy, Harish

Ganapathy, Steven Zheng, HyunJeong Choe, Ágoston Weisz, Tao Zhu, Yifeng Lu, Siddharth Gopal, Jarrod Kahn, Maciej Kula, Jeff Pitman, Rushin Shah, Emanuel Taropa, Majd Al Merey, Martin Baeuml, Zhifeng Chen, Laurent El Shafey, Yujing Zhang, Olcan Sercinoglu, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Gaurav Singh Tomar, Evan Senter, Martin Chadwick, Ilya Kornakov, Nithya Attaluri, Iñaki Iturrate, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Xavier Garcia, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Ravi Addanki, Antoine Miech, Annie Louis, Denis Teplyashin, Geoff Brown, Elliot Catt, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaliy Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo yiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos Campos, Alex

Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Sidharth Mudgal, Romina Stella, Kevin Brooks, Gautam Vasudevan, Chenxi Liu, Mainak Chain, Nivedita Melinkeri, Aaron Cohen, Venus Wang, Kristie Seymore, Sergey Zubkov, Rahul Goel, Summer Yue, Sai Krishnakumaran, Brian Albert, Nate Hurley, Motoki Sano, Anhad Mohananey, Jonah Joughin, Egor Filonov, Tomasz Kępa, Yomna Eldawy, Jiawern Lim, Rahul Rishi, Shirin Badiezadegan, Taylor Bos, Jerry Chang, Sanil Jain, Sri Gayatri Sundara Padmanabhan, Subha Puttagunta, Kalpesh Krishna, Leslie Baker, Norbert Kalb, Vamsi Bedapudi, Adam Kurzrok, Shuntong Lei, Anthony Yu, Oren Litvin, Xiang Zhou, Zhichun Wu, Sam Sobell, Andrea Siciliano, Alan Papir, Robby Neale, Jonas Bragagnolo, Tej Toor, Tina Chen, Valentin Anklin, Feiran Wang, Richie Feng, Milad Gholami, Kevin Ling, Lijuan Liu, Jules Walter, Hamid Moghaddam, Arun Kishore, Jakub Adamek, Tyler Mercado, Jonathan Mallinson, Siddhinita Wandekar, Stephen Cagle, Eran Ofek, Guillermo Garrido, Clemens Lombriser, Maksim Mukha, Botu Sun, Hafeezul Rahman Mohammad, Josip Matak, Yadi Qian, Vikas Peswani, Pawel Janus, Quan Yuan, Leif Schelin, Oana David, Ankur Garg, Yifan He, Oleksii Duzhyi, Anton Älgmyr, Timothée Lottaz, Qi Li, Vikas Yadav, Luyao Xu, Alex Chinien, Rakesh Shivanna, Aleksandr Chuklin, Josie Li, Carrie Spadine, Travis Wolfe, Kareem Mohamed, Subhabrata Das, Zihang Dai, Kyle He, Daniel von Dincklage, Shyam Upadhyay, Akanksha Maurya, Luyan Chi, Sebastian Krause, Khalid Salama, Pam G Rabinovitch, Pavan Kumar Reddy M, Aarush Selvan, Mikhail Dektiarev, Golnaz Ghiasi, Erdem Guven, Himanshu Gupta, Boyi Liu, Deepak Sharma, Idan Heimlich Shtacher, Shachi Paul, Oscar Akerlund, François-Xavier Aubet, Terry Huang, Chen Zhu, Eric Zhu, Elico Teixeira, Matthew Fritze, Francesco Bertolini, Liana-Eleonora Marinescu, Martin Bölle, Dominik Paulus, Khyatti Gupta, Tejasi Latkar, Max Chang, Jason Sanders, Roopa Wilson, Xuewei Wu, Yi-Xuan Tan, Lam Nguyen Thiet, Tulsee Doshi, Sid Lall, Swaroop Mishra, Wanming Chen, Thang Luong, Seth Benjamin, Jasmine Lee, Ewa Andrejczuk, Dominik Rabiej, Vipul Ranjan, Krzysztof Styrc, Pengcheng Yin, Jon Simon, Malcolm Rose Harriott, Mudit Bansal, Alexei Robsky, Geoff Bacon, David Greene, Daniil Mirylenka, Chen Zhou, Obaid Sarvana, Abhimanyu Goyal, Samuel Andermatt, Patrick Siegler, Ben Horn, Assaf Israel, Francesco Pongetti, Chih-Wei "Louis" Chen, Marco Selvatici, Pedro Silva, Kathie Wang, Jackson Tolins, Kelvin Guu, Roey Yogev, Xiaochen Cai, Alessandro Agostini, Maulik Shah, Hung Nguyen, Noah Ó Donnaile, Sébastien Pereira, Linda Friso, Adam Stambler, Adam Kurzrok, Chenkai Kuang, Yan Romanikhin, Mark Geller, ZJ Yan, Kane Jang, Cheng-Chun Lee, Wojciech Fica, Eric Malmi, Qijun Tan, Dan Banica, Daniel Balle, Ryan Pham, Yanping Huang, Diana Avram, Hongzhi Shi, Jasjot Singh, Chris Hidey, Niharika Ahuja, Pranab Saxena, Dan Dooley, Srividya Pranavi Potharaju, Eileen O'Neill, Anand Gokulchandran, Ryan Foley, Kai Zhao, Mike Dusenberry, Yuan Liu, Pulkit Mehta, Ragha Kotikalapudi, Chalence Safranek-Shrader, Andrew Goodman, Joshua Kessinger, Eran Globen, Prateek Kolhar, Chris Gorgolewski, Ali Ibrahim, Yang Song, Ali Eichenbaum, Thomas Brovelli, Sahitya Potluri, Preethi Lahoti, Cip Baetu, Ali Ghorbani, Charles Chen, Andy Crawford, Shalini Pal, Mukund Sridhar, Petru Gurita, Asier Mujika, Igor Petrovski, Pierre-Louis Cedoz, Chenmei Li, Shiyuan Chen, Niccolò Dal Santo, Siddharth Goyal, Jitesh Punjabi, Karthik Kappaganthu, Chester Kwak, Pallavi LV, Sarmishta Velury, Himadri Choudhury, Jamie Hall, Premal Shah, Ricardo Figueira, Matt Thomas, Minjie Lu, Ting Zhou, Chintu Kumar, Thomas Jurdi, Sharat Chikkerur, Yenai Ma, Adams Yu, Soo Kwak, Victor Ähdel, Sujeevan Rajayogam, Travis Choma, Fei Liu, Aditya Barua, Colin Ji, Ji Ho Park, Vincent Hellendoorn, Alex Bailey, Taylan Bilal, Huanjie Zhou, Mehrdad Khatir, Charles Sutton, Wojciech Rzadkowski, Fiona Macintosh, Konstantin Shagin, Paul Medina, Chen Liang, Jinjing Zhou, Pararth Shah, Yingying Bi, Attila Dankovics, Shipra Banga, Sabine Lehmann, Marissa Bredesen, Zifan Lin, John Eric Hoffmann, Jonathan Lai, Raynald Chung, Kai Yang, Nihal Balani, Arthur Bražinskas, Andrei Sozanschi, Matthew Hayes, Héctor Fernández Alcalde, Peter Makarov, Will Chen, Antonio Stella, Liselotte Snijders, Michael Mandl, Ante

Kärrman, Paweł Nowak, Xinyi Wu, Alex Dyck, Krishnan Vaidyanathan, Raghavender R, Jessica Mallet, Mitch Rudominer, Eric Johnston, Sushil Mittal, Akhil Udathu, Janara Christensen, Vishal Verma, Zach Irving, Andreas Santucci, Gamaleldin Elsayed, Elnaz Davoodi, Marin Georgiev, Ian Tenney, Nan Hua, Geoffrey Cideron, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Dylan Scandinaro, Heinrich Jiang, Jasper Snoek, Mukund Sundararajan, Xuezhi Wang, Zack Ontiveros, Itay Karo, Jeremy Cole, Vinu Rajashekhar, Lara Tumeh, Eyal Ben-David, Rishub Jain, Jonathan Uesato, Romina Datta, Oskar Bunyan, Shimu Wu, John Zhang, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Jane Park, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Geoffrey Irving, Edward Loper, Michael Fink, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Ivan Petrychenko, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Evan Palmer, Paul Suganthan, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Ginger Perng, Elena Allica Abellan, Mingyang Zhang, Ishita Dasgupta, Nate Kushman, Ivo Penchev, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnapalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Daniel Andor, Pedro Valenzuela, Minnie Lui, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Ken Franko, Anna Bulanova, Rémi Leblond, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Mark Omernick, Colton Bishop, Rachel Sterneck, Rohan Jain, Jiawei Xia, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Daniel J. Mankowitz, Alex Polozov, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Matthieu Geist, Ser tan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Kathy Wu, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Saaber Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Yeqing Li, Nir Levine, Ariel Stolovich, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Charlie Deck, Hyo Lee, Zonglin Li, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Sho Arora, Christy Koh, Soheil Hassas Yeganeh, Siim Põder, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash Shroff, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Lynette Webb, Sahil Dua, Dong Li, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer

Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Evgenii Eltyshev, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Christof Angermueller, Xiaowei Li, Anoop Sinha, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Denny Zhou, Komal Jalan, Dinghua Li, Blake Hechtman, Parker Schuh, Milad Nasr, Kieran Milan, Vladimir Mikulik, Juliana Franco, Tim Green, Nam Nguyen, Joe Kelley, Aroma Mahendru, Andrea Hu, Joshua Howland, Ben Vargas, Jeffrey Hui, Kshitij Bansal, Vikram Rao, Rakesh Ghiya, Emma Wang, Ke Ye, Jean Michel Sarr, Melanie Moranski Preston, Madeleine Elish, Steve Li, Aakash Kaku, Jigar Gupta, Ice Pasupat, Da-Cheng Juan, Milan Someswar, Tejvi M., Xinyun Chen, Aida Amini, Alex Fabrikant, Eric Chu, Xuanyi Dong, Amruta Muthal, Senaka Buthpitiya, Sarthak Jauhari, Nan Hua, Urvashi Khandelwal, Ayal Hitron, Jie Ren, Larissa Rinaldi, Shahar Drath, Avigail Dabush, Nan-Jiang Jiang, Harshal Godhia, Uli Sachs, Anthony Chen, Yicheng Fan, Hagai Taitelbaum, Hila Noga, Zhuyun Dai, James Wang, Chen Liang, Jenny Hamer, Chun-Sung Ferng, Chenel Elkind, Aviel Atias, Paulina Lee, Vít Listík, Mathias Carlen, Jan van de Kerkhof, Marcin Pikus, Krunoslav Zaher, Paul Müller, Sasha Zykova, Richard Stefanec, Vitaly Gatsko, Christoph Hirnschall, Ashwin Sethi, Xingyu Federico Xu, Chetan Ahuja, Beth Tsai, Anca Stefanoiu, Bo Feng, Keshav Dhandhania, Manish Katyal, Akshay Gupta, Atharva Parulekar, Divya Pitta, Jing Zhao, Vivaan Bhatia, Yashodha Bhavnani, Omar Alhadlaq, Xiaolin Li, Peter Danenberg, Dennis Tu, Alex Pine, Vera Filippova, Abhipso Ghosh, Ben Limonchik, Bhargava Urala, Chaitanya Krishna Lanka, Derik Clive, Yi Sun, Edward Li, Hao Wu, Kevin Hongtongsak, Ianna Li, Kalind Thakkar, Kuanysh Omarov, Kushal Majmundar, Michael Alverson, Michael Kucharski, Mohak Patel, Mudit Jain, Maksim Zabelin, Paolo Pelagatti, Rohan Kohli, Saurabh Kumar, Joseph Kim, Swetha Sankar, Vineet Shah, Lakshmi Ramachandruni, Xiangkai Zeng, Ben Bariach, Laura Weidinger, Tu Vu, Alek Andreev, Antoine He, Kevin Hui, Sheleem Kashem, Amar Subramanya, Sissie Hsiao, Demis Hassabis, Koray Kavukcuoglu, Adam Sadovsky, Quoc Le, Trevor Strohman, Yonghui Wu, Slav Petrov, Jeffrey Dean, and Oriol Vinyals. Gemini: A family of highly capable multimodal models, 2024. URL https://arxiv.org/abs/2312.11805.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/416f9cb3276121c42eebb86352a4354a-Abstract-Conference.html.

Harsh Trivedi, Tushar Khot, Mareike Hartmann, Ruskin Manku, Vinty Dong, Edward Li, Shashank Gupta, Ashish Sabharwal, and Niranjan Balasubramanian. Appworld: A controllable world of apps and people for benchmarking interactive coding agents. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 16022–16076. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.850. URL https://doi.org/10.18653/v1/2024.acl-long.850.

Karthik Valmeekam, Alberto Olmo Hernandez, Sarath Sreedharan, and Subbarao Kambhampati. Large language models still can't plan (A benchmark for llms on planning and reasoning about change). *CoRR*, abs/2206.10498, 2022. doi: 10.48550/ARXIV.2206.10498. URL https://doi.org/10.48550/arXiv.2206.10498.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017. URL http://arxiv.org/abs/1706.03762.

Han Wang, Yongjie Ye, Bingru Li, Yuxiang Nie, Jinghui Lu, Jingqun Tang, Yanjie Wang, and Can Huang. Vision as lora, 2025. URL https://arxiv.org/abs/2503.20680.

Hanlin Wang, Zhan Tong, Kecheng Zheng, Yujun Shen, and Limin Wang. Contextual AD narration with interleaved multimodal sequence. *CoRR*, abs/2403.12922, 2024a. doi: 10.48550/ARXIV.2403.12922. URL https://doi.org/10.48550/arXiv.2403.12922.

Shijie Wang, Qi Zhao, Minh Quan Do, Nakul Agarwal, Kwonjoon Lee, and Chen Sun. Vamos: Versatile action models for video understanding. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XII*, volume 15070 of *Lecture Notes in Computer Science*, pp. 142–160. Springer, 2024b. doi: 10.1007/978-3-031-73254-6_9. URL https://doi.org/10.1007/978-3-031-73254-6_9.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent. *European Conference on Computer Vision (ECCV)*, 2024c.

Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, Sen Xing, Guo Chen, Junting Pan, Jiashuo Yu, Yali Wang, Limin Wang, and Yu Qiao. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022.

Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. Unleashing the emergent cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration, 2024d. URL https://arxiv.org/abs/2307.05300.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa:next phase of question-answering to explaining temporal actions, 2021. URL https://arxiv.org/abs/2105.08276.

Junbin Xiao, Pan Zhou, Angela Yao, Yicong Li, Richang Hong, Shuicheng Yan, and Tat-Seng Chua. Contrastive video question answering via video graph transformer. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(11):13265–13280, 2023. doi: 10.1109/TPAMI.2023.3292266. URL https://doi.org/10.1109/TPAMI.2023.3292266.

Zihui Xue, Kumar Ashutosh, and Kristen Grauman. Learning object state changes in videos: An open-world perspective. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 18493–18503. IEEE, 2024. doi: 10.1109/CVPR52733.2024.01750. URL https://doi.org/10.1109/CVPR52733.2024.01750.

Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 1666–1677. IEEE, 2021. doi: 10.1109/ICCV48922.2021.00171. URL https://doi.org/10.1109/ICCV48922.2021.00171.

Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Zero-shot video question answering via frozen bidirectional language models. In *NeurIPS*, 2022.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. Webshop: Towards scalable real-world web interaction with grounded language agents, 2023a. URL https://arxiv.org/abs/2207.01206.

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 Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023b. URL http://papers.nips.cc/paper_files/paper/2023/hash/271db9922b8d1f4dd7aaef84ed5ac703-Abstract-Conference.html.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models, 2023c. URL https://arxiv.org/abs/2210.03629.

Qinghao Ye, Guohai Xu, Ming Yan, Haiyang Xu, Qi Qian, Ji Zhang, and Fei Huang. Hitea: Hierarchical temporal-aware video-language pre-training. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 15359–15370. IEEE, 2023. doi: 10.1109/ICCV51070.2023.01413. URL https://doi.org/10.1109/ICCV51070.2023.01413.

Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. Self-chained image-language model for video localization and question answering. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/f22a9af8dbb348952b08bd58d4734b50-Abstract-Conference.html.

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. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 21715–21737, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1209. URL https://aclanthology.org/2024.emnlp-main.1209/.

Qi Zhao, Shijie Wang, Ce Zhang, Changcheng Fu, Minh Quan Do, Nakul Agarwal, Kwonjoon Lee, and Chen Sun. Antgpt: Can large language models help long-term action anticipation from videos? In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=Bb21JPnhhr.

Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. Learning video representations from large language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 6586–6597. IEEE, 2023. doi: 10.1109/CVPR52729.2023.00637. URL https://doi.org/10.1109/CVPR52729.2023.00637.

Jie Zhu, Qian Chen, Huaixia Dou, Junhui Li, Lifan Guo, Feng Chen, and Chi Zhang. Dianjin-r1: Evaluating and enhancing financial reasoning in large language models, 2025. URL https://arxiv.org/abs/2504.15716.

A PSEUDOCODE

Algorithm 1: Frame Sampling and Multi-Stage React Planner (MSRP)

/* Step 1: Parameters and Initialization */
  V: Input video, Q, A: Question/answer, k: Top-k selector, $\mathrm{Tool_p}$: Toolset
  Max_Number: Maximum iterations, α: Score threshold
/* Step 2: Frame Sampling */
  $\mathrm{I_s}$: Relevant frames set
  $\mathrm{I} \leftarrow \mathrm{DecodeFrames(V)}$ `// Extract frames from video`
  $\mathrm{I_q} \leftarrow \mathrm{Top_k(Sim(I, Q))}$ `// Select top-k frames for question`
  $\mathrm{I_a} \leftarrow \mathrm{Top_k(Sim(I, A))}$ `// Select top-k frames for answer`
  $\mathrm{I_s} = \mathrm{I_q} \cup \mathrm{I_a}$ `// Combine frames relevant to question and answer`
  $\mathrm{GenerateDescriptions(I_s)}$ `// Generate descriptions using model` $\mathrm{M_c}$
/* Step 3: Multi-Stage Planning */
  $\mathrm{State_{init}} = \emptyset$ `// Initialize planning state`
  while $\neg(\mathrm{Score_e} > \alpha \lor \mathrm{Stopped} \lor \mathrm{Loop} > \mathrm{Max\_Number})$ **do**
    `// Observing Phase //`
    $\mathrm{State_{OBSERVE}} \leftarrow \mathrm{LLM(Prompt_p(State_{init}, I_s, ”OBSERVE”))}$
    `// Thinking Phase //`
    $\mathrm{State_{THINK}} \leftarrow \mathrm{LLM(Prompt_p(State_{OBSERVE}, I_s, ”THINK”))}$
    `// Acting Phase //`
    $\mathrm{Plan} \leftarrow \mathrm{LLM(Prompt_p(State_{THINK}, I_s, ”ACT”, Tool_p))}$
    $\mathrm{ExecutePlan(Plan, Tool_p, M_c)}$
    $\mathrm{Score_e, Reason} \leftarrow \mathrm{Evaluation(I_s)}$
    $\mathrm{State_{init}} \leftarrow \mathrm{State_{THINK}}$
end
`// * Evaluation Function * //`
  $\mathrm{Evaluation(I_s)}$:
    $\mathrm{Scores} \leftarrow \mathrm{AssessClarity(I_s, 60\%) + AssessAlignment(I_s, 40\%)}$
    return $\mathrm{Scores, Feedback}$ `// Evaluation result`
`// * Final Output * //`
  return $\mathrm{I_{IA}}$

Algorithm 2: Multi-Agent Debate Refinement (MADR)

/* Step 1: Initial Answer Generation */
$\mathrm{Answer_{initial}} \leftarrow \mathrm{LLM(Prompt_{IA}(M_c(I_{IA}), Question, Options))}$ `// Generate answer with reasoning`
/* Step 2: Multi-Agent Debate */
  Define roles: $\mathrm{Affirmative, Opposition, Judge}$
  $\mathrm{Statement} \leftarrow \mathrm{Answer_{initial}}$ `// Initialize debate input`
while $\mathrm{Cond_{db}}$ **do**
    $\mathrm{Affirmative\ Statement} \leftarrow \mathrm{LLM(Prompt_{db}(Task_{Affirmative}, Statement, Tool_{db}))}$
    `// Support answer //`
    $\mathrm{Opposition\ Statement} \leftarrow \mathrm{LLM(Prompt_{db}(Task_{Opposition}, Statement, Tool_{db}))}$
    `// Challenge answer //`
    $\mathrm{Decision} \leftarrow \mathrm{Judge(Assess(Affirmative\ Statement, Opposition\ Statement))}$
    `// Judge evaluates arguments //`
    $\mathrm{Round\ Number} \leftarrow \mathrm{Round\ Number} + 1$ Increment round
end
`// * Debate Termination Conditions * //`
$\mathrm{Cond_{db}} \leftarrow \mathrm{Agreement} \lor \mathrm{Win} \lor (\mathrm{Round\ Number} > \mathrm{Max\ Number})$
  Agreement: Both sides reach consensus
  Win: One side prevails over the other
  Round Number: Tracks debate iterations
`// * Final Output * //`
  return $\mathrm{Answer_{updated}}$

Table 4: **Performance comparison across various scales of LLM (VideoAgent).** Scale indicates the parameter size of the LLM, all of which are from the Qwen 2.5 series. Subset refers to the validation set of EgoSchema. All videos in EgoSchema were concatenated into a single approximately **28-hour-long** video to evaluate the performance on all questions in the Subset, Long (28h).

Scale	EgoSchema		Next QA (Val)				Next QA (ATP-Hard)		
	Long (28h)	Subset	All	Causual	Temporal	Descriptive	All	Causual	Temporal
72 B	33.8	56.8	71.0	72.1	63.8	82.6	62.3	65.7	57.5
32 B	33.0	51.4	69.0	70.5	61.1	80.2	60.0	64.2	53.9
14 B	33.2	51.2	69.6	71.1	62.3	78.8	61.5	65.9	54.9
7 B	31.6	47.8	66.4	67.0	60.7	75.3	58.2	61.1	53.7
3 B	29.4	48.0	55.0	54.6	49.8	66.4	46.4	48.7	43.2
1.5 B	23.4	32.8	49.1	47.7	47.7	61.6	42.1	42.8	40.8

B Datasets and Metrics

EgoSchema Mangalam et al. (2023). EgoSchema is a standard benchmark designed for long-video understanding. It consists of approximately 5,000 real-world video samples, covering various human activities such as cooking, crafting, and sports. Each video is unedited, captured from a first-person perspective, and has a minimum duration of 3 minutes. The dataset's total duration exceeds 250 hours, with an average video length of 3 minutes and 17 seconds. For each video, a multiple-choice question with five candidate options is provided. EgoSchema is specifically designed to evaluate long-video understanding in zero-shot settings. The dataset is organized into two parts: a subset and the full set. The subset contains 500 QA tasks with corresponding annotations, while the full set includes all video. Performance on the full set requires submitting results to the official server.

Next QA Xiao et al. (2021). Next QA is a video understanding benchmark designed to evaluate a model's capabilities in causal reasoning, temporal analysis, and scene understanding. The dataset consists of 5,440 videos, with an average duration of 44 seconds. The content of the videos includes scenes of daily activities and interpersonal interactions. A total of 52,000 question-answer (QA) pairs are annotated for these videos, with causal reasoning, temporal analysis, and scene understanding accounting for 48%, 29%, and 23% of the QA tasks, respectively. Consistent with prior work, we conduct our analysis only on a subset of Next QA. In addition, we further partitioned this subset into a more challenging ATP-hard subset Buch et al. (2022). QA tasks in the ATP-hard subset cannot be resolved using single-frame images and are designed to assess long-term reasoning ability.

C VideoAgent Performance based on Qwen

In a direct comparison against VideoAgent on the same Qwen series models, RIVAL shows a significant performance advantage (Figure 4). For instance, on Qwen 2.5 72B, we achieved a performance improvement of 3.4. Across other scales, RIVAL maintains a consistent lead. Even on Qwen 2.5 1.5B, RIVAL still outperforms VideoAgent by 4.5 in accuracy. On more challenging benchmarks, such as EgoSchema's evaluation dataset, the performance gap between RIVAL and VideoAgent further widens. For example, on Qwen 2.5 72B, RIVAL achieved a 10-point improvement in accuracy. In the ultra-long video evaluation scenario (Long), our method significantly surpassed VideoAgent in performance. When facing nearly 28 hours of video input, VideoAgent's performance noticeably deteriorates; for Qwen 2.5 1.5B, its accuracy drops to 23.4, barely above random selection. In contrast, our RIVAL manages to interpret the 28-hour video with only minor performance degradation. On Qwen 2.5 1.5B, RIVAL still achieved an accuracy of 33.8, outperforming VideoAgent by nearly 10%.

D Ablation Study

Parameter Ablation Study for RIVAL. In Figure 4, we present the parameter ablation experiments for RIVAL. Five hyperparameters are analyzed, including the Top-k value for the initial search, the threshold for terminating loops in MSRP, the maximum number of iterations, as well as the threshold and number of steps for initiating the debate process. It is important to highlight that

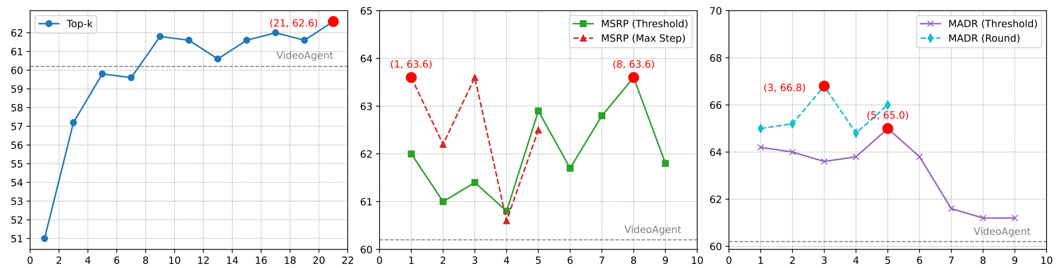

Figure 4: **Parameter Ablation Experiments on EgoSchema Benchmark.** From left to right, the figures depict the parameters for initial information retrieval, the threshold for terminating loops in MSRP, and the maximum number of iterations. The final figure illustrates the threshold for answer refinement and the maximum number of debate rounds. In each figure, the best performance achieved by the SOTA GPT-4-based VideoAgent on the EgoSchema is indicated by gray dashed lines.

RIVAL demonstrates robustness and effectiveness. After completing the initial information search, the remaining parameter settings are designed to explore the upper bounds of RIVAL's performance. As shown in the second and third subfigures, RIVAL is capable of outperforming the previous state-of-the-art GPT-4-based VideoAgent even under relatively unfavorable parameter settings.

Top-k Value Analysis. The impact of the top-k parameter on model performance is illustrated in the first subfigure of Figure 4. During the initial information collection stage, both the question and answer are used to retrieve corresponding frames, which are subsequently described using a Caption model. Intuitively, retrieving more frames provides additional information, offering greater support for subsequent answer generation. However, excessive information may lead to redundancy, negatively affecting the model's decision-making. As shown in the first subfigure of Figure 4, initially increasing the top-k value leads to a significant improvement in model performance, with accuracy rising sharply from 51% at top-k = 1 to nearly 62% at top-k = 9. Beyond this point, while performance continues to improve, it becomes slower and exhibits slight fluctuations until top-k reaches 21. Therefore, to avoid introducing excessive irrelevant information and optimize performance, we select top-k = 21 as our final setting, at which the model achieves an accuracy of 62.6%.

Threshold for MSRP Loops. The second subfigure of Figure 4 illustrates the impact of the loop threshold in MSRP, represented by the green solid line. After the initial information is retrieved, the planner and evaluator collaborate to gather additional information while filtering out irrelevant data. The loop threshold determines whether the iterative process should continue. Intuitively, applying the loop to all available information is expected to yield better results, which aligns with our experimental findings. As shown in Figure 4, overall model performance improves as the loop threshold increases. These results further validate the rationale behind the collaborative loop between the planner and evaluator. Based on this observation, we select a loop threshold of 8.

Iteration Steps in MSRP Loops. In the second subfigure of Figure 4, we use a red dashed line to illustrate the impact of the number of iteration steps on model performance. A greater number of iteration steps results in longer contextual content. However, since we use the 72B network as the baseline model, its ability to support extended contexts is limited compared to larger models, such as GPT 4. Consequently, model performance gradually decreases as the number of iterations increases. To mitigate issues arising from excessively long contextual content, we set maximum number to 1.

Threshold for Debate. The third subfigure of Figure 4 shows the trend of model performance as the debate threshold varies, represented by the purple solid line. During the debate process, we observed that applying debate to all questions with different scores had varying impacts on performance. Specifically, debating low-scoring questions was less effective compared to setting a higher threshold and only debating high-scoring ones. For instance, using a threshold of 1, where answers with evaluator scores greater than 1 are debated, was found to be less effective than setting the threshold to 5. Through visualization, we discovered that low-scoring information often fails to perfectly capture the correct corresponding options, making it prone to being dominated by the opposing side during the debate process. Therefore, we set the threshold to 5, ensuring that only results with scores greater than 5 are fine-tuned through debate.

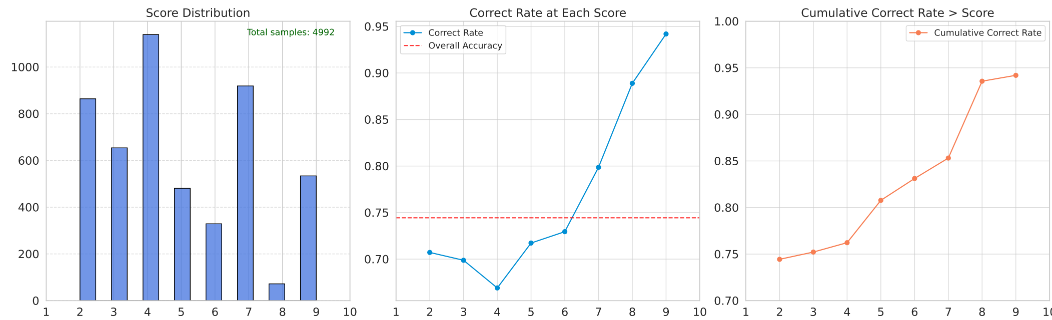

Figure 5: **Score Analysis on the Next QA Benchmark.** From left to right, the figure presents the distribution of scores after iterative processing, the accuracy corresponding to each score value, and the accuracy for scores greater than the current score. For example, the accuracy within the range (2, 10]. Overall, the accuracy of RIVAL improves as the evaluation score increases.

Rounds of Debate. The third subfigure of Figure 4 illustrates the trend of model performance as the number of debate rounds increases, represented by the dashed line. The performance of RIVAL improves with an increasing number of debate rounds. When the number of rounds is set to 3, the performance reaches its peak value of 66.8, which aligns with the results we reported in the experimental section. Therefore, we hypothesize that increasing the number of debate rounds could further enhance the performance of RIVAL, as long as the model's ability to comprehend contextual information remains uncompromised. For simplicity, we set the number of debate rounds to 3.

E Evaluation Analysis

Whether guiding decision-making for the planner in MSRP loops or using thresholds in the MADR model to filter low-scoring samples, the effectiveness of our method, RIVAL, is fundamentally dependent on the accuracy of the evaluation scores. To validate this dependence, we conducted a series of data analyses on scores generated by the evaluator, as shown in Figure 5. Due to the limited sample size of the EgoSchema benchmark, which contains only 500 samples, the representativeness of the data is constrained, leading to larger deviations in score distribution. To further ensure the reliability of our analysis, we extended the experiments to the Next QA benchmark, which includes nearly 5,000 samples and offers broader representativeness.

The analysis of this dataset reveals that a larger proportion of questions are clustered in the lower score ranges, where accuracy is typically lower. Specifically, when the evaluation score equals 2, the prediction accuracy of RIVAL is approximately 71%, whereas at a score of 9, the accuracy rises to nearly 95%. This finding indicates a strong positive correlation between the evaluator's scoring capabilities and the accuracy of RIVAL predictions. From an interval-based perspective, this trend is particularly prominent: in lower score intervals (e.g., a score of 2), the accuracy is around 75%, while in higher score intervals (e.g., a score of 9), the accuracy quickly climbs to nearly 95%.

The collaborative workflow of RIVAL, which integrates the evaluator and planner, proves to be both reasonable and efficient. By decomposing the overall system capability into multiple independent but mutually cooperative components, we are able to reduce the context length while leveraging the advantages of multi-agent collaboration, thereby unlocking the potential of collective intelligence.

F Prompt Design

Prompt Design in Evaluator (Figure 6). Figure 6 illustrates the prompt design in the Evaluator. In the prompt, we specify the evaluation criteria for the LLM and assign weights of 60% and 40%, respectively. For each scoring level, evaluation standards are explicitly defined. Furthermore, good and bad examples are provided to users within the prompt. The output format is restricted to JSON, and the model is required to include explanations for its outputs.

```
You are an AI assistant tasked with evaluating the accuracy and completeness of provided descriptions answering a given question. Focus on identifying the correctness, relevance, and clarity of the descriptions, as well as whether they have enough detail to fully answer the question. You must assign scores strictly between 1-10 (10 = best).

  #### **Scoring Guidelines**
  1. **Accuracy (60% weight):** How well the description matches the question.
     - **1-2:** Irrelevant, incorrect, or misleading.
     - **3-4:** Weak alignment, unclear or partially wrong.
     - **5-6:** Moderately accurate but incomplete or noisy.
     - **7-9:** Mostly accurate with minor imperfections.
     - **10:** Perfectly correct and directly answers the question.

  2. **Completeness (40% weight):** Does the description provide enough detail to fully answer the question?
     - **1-2:** Extremely vague or lacking critical details.
     - **3-4:** Missing important aspects, limited information.
     - **5-6:** Partially complete, with notable omissions.
     - **7-9:** Mostly complete, minor missing details.
     - **10:** Fully detailed with all relevant context included.

  ### **Your Output Format**
     Respond strictly in JSON format.
```

System Prompt

```
Task Context:
  - Description List Format (For Reference):
  A list containing segment dictionaries with:
     - "segment_id": string identifier (e.g., "1")
     - "frame": video time range (e.g., "0:05-0:10")
     - "description": text with #C/#O tags

  - Action Tags:
  #C = Camera wearer's action\n#O = Others' actions

  - Required Output Format:
  JSON object with:
     - "score": number between 1-10
     - "reason": scoring rationale

  - Example of a correct response:
  {
     "score": 8,
     "reason": "The descriptions partially answer the question but lack some key details."
  }

  Input Data:
  - Historical Evaluations: {eval_history}
  - Current Description List: {captions}
  - Question: {Question}
  - Options: {Optoins}
  Requirement:
  1. Evaluate the current description list's accuracy in answering the question using the provided Scoring Guidelines.
  2. Before assigning the final score, ALWAYS perform a comparison between the current description and the following examples:

     **Positive Example:** "#C picks up a red pencil from the table (frame: 0:05-0:10)"
     - This description is clear, detailed, and fully answers the question with subject, object, action, and time frame. set score to 8

     **Negative Example:** "#C does something (frame: 0:05-0:10)"
     - This description is vague, lacks specific details, and fails to directly address the question. set score to 2

  3. Based on this comparison, justify whether the current description aligns closer to the positive or negative example and
how this influenced your scoring. Your reasoning in the `"reason"` field MUST include reference to this comparison.
```

User Prompt

Figure 6: **Prompt Design in Evaluator.** Variable parameters are treated as input variables and are highlighted in the figure for clarity. These parameters include evaluation history, existing descriptive information, and the questions and options of the QA tasks.

Prompt Design in MSRP (Figure 7). Figure 7 presents the prompt design within MSRP. In MSRP, we take the output from the Evaluator as the evaluation direction. This process unfolds in three steps, each corresponding to a distinct stage of the ReAct paradigm. At each stage, the LLM is required to complete the task specified for that stage. For example, in Stage 2, the LLM is prompted to reflect on how to improve the current descriptive information based on prior observations, and this reflection is provided as a strategy for the next stage. In the final stage, the Agent is granted the ability to invoke tools and instructed to implement the previous strategy through function calls. These designs enforce the LLM to execute plan generation tasks strictly in alignment with the ReAct paradigm.

```
STEP 1: OBSERVE
   Carefully observe the provided input data, including the video description list, evaluation results, and previous plan results.
   Summarize key facts and formulate insights that will be useful in generating a plan.

STEP 2: THINK
   Based on the observed facts from the previous step, generate actionable strategies for solving the task.
   Include reasoning behind your strategies.

STEP 3: ACT
   Implement strategies developed in the previous phase by making function calls that optimize the task outcome
```

System Prompt

```
STEP 1: OBSERVE
Task Background and Context:
   - You are an AI assistant tasked with generating an intelligent plan based on video descriptions, previous results, evaluations, and user-defined questions.
   - Input consists of video frames ({num_frames} @1 fps), a description list with actions (#C = Action by camera wearer, #O = Action by others), previous planning results, and evaluation metrics. Follow the "Observe-Think-Act" methodology to generate actionable plans.
   Your role:
   - - At each step, ensure your reasoning and actions align with maximizing task outcomes.
      Observation Phase Input:
      - Description List Format:
         Structure of Description List:
            A JSON list where each item includes:
            - "segment_id": string (e.g., "1")
            - "frame": frame range (e.g., "100-200")
            - "description": natural language text with #C/#O action tags.
      - Current Description List: {captions}
      - Evaluation Results: {eval_result}
      - Previous Plan Result: {plan_history}
      - Question: {question}
      - Options: {option}

      Requirement:
      - Summarize key observations related to the given inputs.

STEP 2: THINK
Task Background and Context:
   - You are an AI assistant tasked with generating an intelligent plan based on video descriptions, previous results, evaluations, and user-defined questions.
   - Input consists of video frames ({num_frames} @1 fps), a description list with actions (#C = Action by camera wearer, #O = Action by others), previous planning results, and evaluation metrics.
      Thinking Phase Input:
      - Observed Facts: {Step1_output}
      - Question: {question}
      - Options: {option}

      Requirement:
      - Develop strategies and reasoning based on the observations.

STEP 3: ACT
Task Background and Context:
   - You are an AI assistant tasked with generating an intelligent plan based on video descriptions, previous results, evaluations, and user-defined questions.
   - Input consists of video frames ({num_frames} @1 fps), a description list with actions (#C = Action by camera wearer, #O = Action by others), previous planning results, and evaluation metrics.
      Action Phase Input:
      - Proposed Strategies: {Step2_output}
      - Available Function Calls: {available functions}

      Requirement:
      - Execute actions using available function calls based on the strategies.
```

User Prompt

Figure 7: **Prompt Design in MSRP.** In the MSRP framework, variable parameters are treated as input variables and highlighted in the figure for clarity. These variable parameters include evaluation history, existing descriptive information, and the questions and options associated with QA tasks.

Prompt Design for Initial Answer Generation and Debate Roles of Proponents and Opponents (Figure 8). For simplicity, during the initial answer generation process, we use prompts provided by the VideoAgent that include only basic background information, such as the number of video frames. The LLM generates responses directly while being required to provide reasoning for its predictions. In the debate prompt design, we employ a similar structure for both proponents and opponents. Their prompts explicitly outline their respective tasks: the proponent searches for information to support the provided answer, while the opponent seeks information to refute it. Furthermore, both sides are granted the ability to use tools. During each turn of the debate, they are allowed to invoke a tool once. To streamline the process, both sides utilize the same user template.

System Prompt
You are a helpful assistant designed to output JSON.
User Prompt
Given a video that has {num_frames} frames, the frames are decoded at 1 fps. Given the following descriptions of the sampled frames in the video:
{caption}
#C to denote the sentence is an action done by the camera wearer (the person who recorded the video while wearing a camera on their head).
#O to denote that the sentence is an action done by someone other than the camera wearer.
Please answer the following question:
```
{question}
```
Please think carefully and write the best answer index in Json format {'answer': 'xxx', 'reason': 'xxx'}. Note that only one answer is returned for the question, and ** you must select one answer index from the candidates (0/1/2/3/4)**.

You are a debater representing the affirmative side, defending the given answer. Your task:

1. Use descriptors or known facts to justify the answer.
2. Refute negative arguments by identifying flaws and assumptions.
3. Strengthen your stance with expanded explanations if necessary.
4. Default to supporting the provided answer unless evidence strongly contradicts it.
5. For each responce, you have one chance to call the function to collect more information to support you answer.

Output in JSON:
{
'answer': the option you defend,
'evidence': supporting evidence (≤50 words).
}

System Prompt

You are a highly analytical debater representing the opposing side. Your task:

1. Aggressively challenge the given answer by identifying logical flaws, overgeneralizations, contradictions, or gaps in evidence.
2. Critically analyze even subtle weaknesses in affirmative reasoning.
3. If proposing an alternative answer, provide compelling evidence and reasoning to justify it.
4. Avoid speculative or baseless challenges; all objections must be rooted in logic, context, or provided evidence.
5. When no flaws exist and evidence is sound, reluctantly agree with the affirmative side.
6. For each responce, you have one chance to call the function to collect more information to support you answer.

Output in JSON:
{
"answer": 0/1/2/3/4, the option you believe is correct (mandatory; no null or none allowed),
"evidence": Clear and concise supporting evidence, including detailed reasoning behind your objection or agreement (≤50 words).
}

System Prompt

Task Context: a video that has {num_frames} frames, the frames are decoded at 1 fps. Frames are sampled from the video, and their descriptions are extracted: {caption}. #C to denote the sentence is an action done by the camera wearer (the person who recorded the video while wearing a camera on their head). #O to denote that the sentence is an action done by someone other than the camera wearer.

Question: {question}.

Given init answer: {init_answer}.

User Prompt

Figure 8: **Prompt Design for the Initial Answer Generation and Roles of Proponents and Opponents in the Debate.** The variable parameters are treated as input variables and are highlighted in the figure for clarity. These parameters include existing descriptive information and the questions and options of the QA tasks. To prevent exceeding the context window, we constrain their evidence to within 50 words.

Prompt Design for the Judge (Figure 9). Figure 9 illustrates the prompt design for the judge. The judge performs two main functions: determining whether the debate process should continue and generating the final answer. For process determination, the judge simply evaluates whether the proponent and opponent have reached a consensus. For generating the final answer, the judge collects the debate's historical information and formulates the final answer accordingly.

G The Use of Large Language Models

In this work, we utilized Large Language Models (LLMs) exclusively for polishing the text of the manuscript.

```
System Prompt (in Debate)
You are an impartial moderator tasked with evaluating a debate between two sides: Affirmative and Negative. Your task:
    1. Objectively assess the strength of reasoning, evidence, and rebuttal effectiveness from both sides, without default bias.
    2. Favor Affirmative if its arguments hold up under scrutiny and Negative fails to provide strong counterarguments.
    3. Favor Negative if its objections are well-reasoned, logical, and backed by compelling evidence that decisively undermines Affirmative's answer.
    4. If Negative fails to select an answer in the 0-4 range, Affirmative wins by default.
    5. Always provide a clear and neutral summary of your reasoning for the decision.
    Output in JSON:
    {
      "continue": True/False,
      "choice": "Affirmative"/"Negative",
      "reason": "Neutral and concise summary of the decision-making rationale."
    }
User Prompt (in Debate)
Q & A Mission Statement: {}, Affirmative: {} Negative: {}
```

```
System Prompt (Give Answer)
You are an impartial judge tasked with determining the most accurate answer based on arguments presented by both sides. Your task:
    1. Select the option supported by the strongest evidence, reasoning, and overall argument quality.
    2. **Give priority to consensus when both sides recommend the same answer; only evaluate the quality of evidence and arguments when their answers differ.**
    3. Ensure your judgment is based purely on logical and factual merit, avoiding any bias or assumptions.
    4. Provide a clear and neutral summary of your reasoning for the chosen answer.
    Output in JSON:
      {
        "final_answer": 0/1/2/3/4,
        "reason": "Neutral and concise summary of judgment rationale."
      }
User Prompt (Give Answer)
Based on the debate, select the most accurate answer.
  Output:
  {
    'final_answer': 0/1/2/3/4,
    'reason': 'Reason for final choice.'
  }
```

Figure 9: **Prompt Design for Debate Roles.** The variable parameters are treated as input variables and are highlighted in the figure for clarity. These parameters include the questions, options of the QA tasks, and the dialogue history of both sides in the debate. Note that the decision of whether to continue the debate and the subsequent generation of the final answer are separated processes. However, as both serve summarization functions, we classify them under the same role.