

DIVE: Deep-search Iterative Video Exploration

A Technical Report for the CVRR Challenge at CVPR 2025

Umihiro Kamoto Tatsuya Ishibashi Noriyuki Kugo
Panasonic Connect Co., Ltd.

{kamoto.umihiro, ishibashi.tatsuya001, kugou.noriyuki}@jp.panasonic.com

Abstract

*In this report, we present the winning solution that achieved the 1st place in the Complex Video Reasoning & Robustness Evaluation Challenge 2025. This challenge evaluates the ability to generate accurate natural language answers to questions about diverse, real-world video clips. It uses the Complex Video Reasoning and Robustness Evaluation Suite (CVRR-ES) benchmark, which consists of 214 unique videos and 2,400 question-answer pairs spanning 11 categories. Our method, **DIVE (Deep-search Iterative Video Exploration)**, adopts an iterative reasoning approach, in which each input question is semantically decomposed and solved through stepwise reasoning and progressive inference. This enables our system to provide highly accurate and contextually appropriate answers to even the most complex queries. Applied to the CVRR-ES benchmark, our approach achieves 81.44% accuracy on the test set, securing the top position among all participants. This report details our methodology and provides a comprehensive analysis of the experimental results, demonstrating the effectiveness of our iterative reasoning framework in achieving robust video question answering. The code is available at <https://github.com/PanasonicConnect/DIVE>.*

1. Introduction

Video understanding is a critical research challenge in both computer vision and natural language processing. In recent years, the Video Question Answering (VQA) task, which requires models to answer a question about video clips, has emerged as a comprehensive benchmark for assessing video understanding capabilities. Approaches to VQA have become increasingly diverse, with extensive research on end-to-end methods [5–7, 11, 14, 17] and approaches that convert video content into image captions and leverage large language models (LLMs) to generate answers [9, 15, 16, 18, 19]. More recently, methods that

incorporate AI agents have also gained attention. For example, some approaches dynamically generate specialized agents based on the question, while others employ agents to select important video frames [1, 3, 4, 12, 13, 20]. Iterative reasoning approaches, such as OpenAI’s DeepResearch [10], which employ multi-step inference strategies, have advanced LLMs by enabling a deeper and more robust understanding of complex questions. However, these techniques are still rarely applied to VQA, especially for video understanding. As a result, many existing VQA methods answer questions in a straightforward manner, often overlooking the underlying intent and producing less contextually appropriate responses. To address these gaps, we propose incorporating intent estimation and iterative reasoning to further improve video question answering.

In this paper, we present DIVE (Deep-search Iterative Video Exploration), a novel framework that combines semantic decomposition, intent estimation, and iterative inference to generate accurate answers to questions about given video clips. Specifically, DIVE breaks down each question into sub-questions based on the underlying intent, and solves them through an iterative process to generate contextually appropriate and precise answers. Additionally, we propose an object-centric video summarization method. This method uses object detection techniques to create summaries focused on the appearance and spatio-temporal transitions of key objects in each scene. This leads to improved overall performance in video understanding. The main contributions of this work are as follows:

- We propose DIVE, a new framework that answers complex video questions by breaking them down into sub-questions and solving them through an iterative reasoning process.
- We incorporate intent estimation into the VQA task to enable answers that better capture the underlying intent behind each question.
- We develop a novel object-centric video summarization method that enhances video comprehension by capturing key object transitions.

Our results demonstrate that our method effectively and

robustly addresses the complex queries featured in the CVRR-ES benchmark.

2. Methodology

Figure 1 presents the overall architecture of our system, which incrementally and iteratively enhances video understanding and answer accuracy through a six-step process: (1) intent estimation, (2) question breakdown into sub-questions, (3) agent-based answering of sub-questions, (4) refinement of remaining sub-questions, (5) loop continuation judgment, and (6) final answer generation by integrating all answers to sub-questions. The following subsections describe each step in detail.

2.1. Step 1: Intent Estimation

In Step 1, intent estimation is performed on the input question. The system interprets what the question is fundamentally asking by considering both the question text and the video context. Specifically, a video summary is used as the key visual information from the video. The question and this video summary are then combined to estimate and describe in detail the underlying intent of the question.

2.2. Step 2: Question Breakdown

In Step 2, the question is decomposed into multiple sub-questions. At this step, both the question text and the video summary information are provided as input, and relevant sub-questions are generated based on the original question. This decomposition enables the system to focus on specific parts or aspects of the video. Furthermore, information about the video analysis tools to be used by the agent in Step 3 is also considered when generating the sub-questions. By tailoring the sub-questions to match the analytical capabilities of the subsequent agent, the system enables more effective and targeted video analysis.

2.3. Step 3: Question Answering

In Step 3, an agent is employed to answer the sub-question with the highest priority in each loop. The remaining sub-questions are retained for further refinement in the next step. For each sub-question, the agent utilizes the following dedicated video analysis tools to produce appropriate responses:

- **Gemini 2.5 Pro Tool:** Analyzes one frame per second along with the corresponding audio. This tool excels in temporal reasoning, recognizing audio cues. Because it processes the entire video at a consistent temporal resolution, it is particularly useful for achieving a comprehensive understanding of the overall video content.
- **GPT-4.1 Tool:** Based on the sub-question and video summary information, this tool identifies key temporal

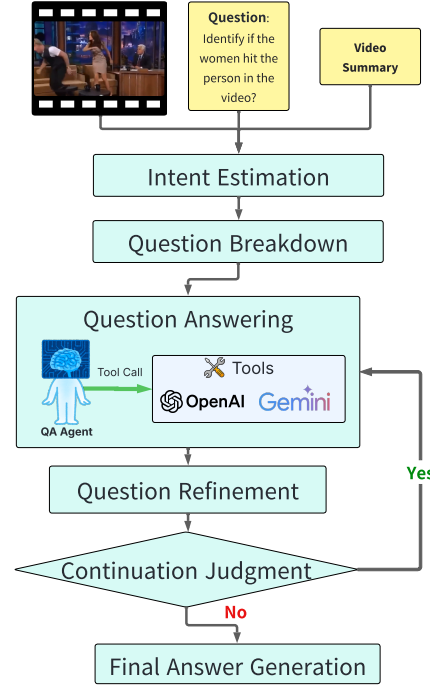


Figure 1. Overall architecture of DIVE: Achieving robust and accurate VideoQA by iteratively solving sub-questions decomposed by the question breakdown module.

segments of interest, samples frames from those segments, and performs detailed analysis using GPT-4.1. By selecting and analyzing 8-16 frames that are most relevant to the sub-question, this tool enables detailed analysis of specific temporal segments within the video. By leveraging these tools, the agent can generate optimal answers for each sub-question.

2.4. Step 4: Question Refinement

In Step 4, the remaining sub-questions are refined and adjusted as needed based on the agent’s responses. For example, if a sub-question regarding the existence of an object receives a negative answer, additional follow-up sub-questions may be generated to reconfirm the result. In this way, the sub-questions are flexibly modified or supplemented according to the situation, enabling the confirmation of necessary information and more precise video understanding through optimized question flow and content.

2.5. Step 5: Continuation Judgment

In Step 5, the system decides whether to continue the loop. The need for additional information is assessed based on the agent’s responses and the current set of sub-questions. If additional information is required to confidently answer the original question, the process returns to Step 3. In contrast,

when sufficient information has been gathered, particularly for straightforward questions, the system terminates the loop to avoid “overthinking” and maintain the quality of the answers. This iterative procedure enables continuous and adaptive information gathering, facilitating a deeper understanding of the video while avoiding unnecessary analysis.

2.6. Step 6: Final Answer Generation

In Step 6, the final answer is generated. The system synthesizes the agent’s responses to the sub-questions together with the video summary to produce a comprehensive answer to the original question. By integrating these sources of information, the system provides a contextually accurate and precise response.

2.7. Video Summarization Using Object Detection

Our video summarization process leverages object detection results to generate context-aware summaries for each video. The procedure consists of three main steps:

1. **Object Label Extraction:** 32 evenly spaced frames are sampled from each video, and object labels are extracted from these frames using GPT-4.1.
2. **Object Detection:** Using the extracted object labels as prompts, we perform object detection on all video frames using Grounding DINO [8], capturing the presence and spatial locations of these objects throughout the video.
3. **Video Summary Generation:** A video summary is generated using GPT-4.1 by integrating the sampled frames and the object detection results, describing key objects, their transitions, and interactions throughout the video.

This approach enables the creation of video summaries that reflect both global context and fine-grained object-level information, thereby supporting more accurate video understanding and downstream tasks.

3. Experiments

In this section, we present the evaluation of our proposed DIVE on the CVRR-ES benchmark. We first provide implementation details of our DIVE architecture in Section 3.1. In Section 3.2, we compare our method with existing approaches. We then conduct an in-depth ablation study in Section 3.3, where we analyze the impact of each component and include detailed case studies to illustrate the effectiveness and characteristics of our approach.

3.1. Implementation Details

Our proposed method, DIVE, was implemented using the open-source LangGraph library ¹. For all steps, we

¹LangGraph, <https://www.langchain.com/langgraph>

Table 1. Comparison with EvalAI leaderboard entries and paper-reported baselines on the CVRR-ES validation set.

Method	Acc. (%)
Baseline (GPT4V) [2]	70.78
Baseline (GPT-4o) [2]	75.03
FRI	53
Host_6403_Team	63
NJUST_KMG	85
PCIEgogogo	88
DIVE	91.55

Table 2. Comparison with the leaderboard entries announced by the organizing committee on the CVRR-ES test set.

Method	Acc. (%)
aaa_vlm	65.23
PCIEgo	73.69
PCIEgogogo	74.95
PCIE	75.32
love_liang	77.84
NJUST_KMG	78.02
DIVE	81.44

used an agent and LLM based on OpenAI’s GPT-4.1 ² (version gpt-4.1-2025-04-14). For the tools invoked by the agent during Step 3 (Question Answering), we utilized Gemini 2.5 Pro, provided by Google Vertex AI ³ (version gemini-2.5-pro-preview-05-06), as the Gemini 2.5 Pro Tool for video analysis. For frame-level image analysis, we employed OpenAI’s GPT-4.1 within the GPT-4.1 Tool. The temperature parameter for the LLMs used in each tool was set to 1.0, and to 0.0 for the agent itself. Additionally, the maximum number of reasoning steps in DIVE was set to 25.

3.2. Main Results

Table 1 summarizes the performance of existing methods and participants on EvalAI public leaderboard for the CVRR-ES validation set, while Table 2 presents the results for the test set from the leaderboard announced by the organizing committee. As shown in Table 1, our proposed method achieves an accuracy of 91.55% on the validation set, significantly surpassing both the highest previously reported score of 75.03% from the CVRR-ES paper [2] and the scores of all other teams on the public leaderboard. Likewise, as shown in Table 2, our method achieves the best performance on the test set with an accuracy of 81.44%, again outperforming all competing approaches on the leaderboard. These results demonstrate the superior

²OpenAI, <https://openai.com>

³Vertex AI, <https://cloud.google.com/vertex-ai>

Table 3. Ablation study on CVRR-ES validation set. GPT-4.1 refers to direct API use; QA Agent employs both GPT-4.1 and Gemini 2.5 Pro tools. Checkmarks (✓) indicate enabled components, and values in parentheses show incremental accuracy gains with each added component.

QA Method	Breakdown	Intent	Summary	Acc. (%)
GPT-4.1				81.49
GPT-4.1	✓			85.83 (+4.34)
QA Agent	✓	✓		87.58 (+1.75)
QA Agent	✓	✓	✓	88.00 (+0.42)

robustness and accuracy of our approach on the CVRR-ES benchmark compared to existing methods.

3.3. Ablation Study

We conduct an ablation study to evaluate the effectiveness of key components in our proposed method. We progressively integrate individual modules into a base pipeline and assess the contribution of each component to overall performance. Specifically, we investigate three key components: agent-based QA for answering sub-questions, intent estimation in Step 1, and video summarization using object detection for Step 2 (Question Breakdown). All experiments are conducted on the CVRR-ES validation set following the evaluation protocol of [2].

Table 3 shows the results of our ablation study on the CVRR-ES validation set. Starting from the GPT-4.1 baseline, adding the Question Breakdown module increases accuracy from 81.49% to 85.83%. Incorporating Intent Estimation and QA Agent further boosts performance to 87.58%, and the addition of Video Summarization yields the best result of 88.00%. These findings highlight the incremental contribution of each component to the overall effectiveness of our approach.

In addition to quantitative results, Figure 2 presents a representative example from the CVRR-ES benchmark where both Gemini 2.5 Pro and GPT-4.1 fail to produce the correct answer. In contrast, our method, DIVE, successfully solves the problem by breaking down the complex question into sub-questions and applying iterative reasoning. This qualitative result further underscores the strength of our modular approach, particularly the question breakdown and iterative reasoning components, in addressing complex queries that require deeper understanding and iterative inference.

4. Conclusion

In this report, we have proposed and evaluated a robust video question answering method that achieved 1st place in the Complex Video Reasoning & Robustness Evaluation Challenge 2025. Our approach combines iterative

Sub-Q1: Is there a boy and a girl present in the video?

Sub-A1: Based on the visual cues in the frames, there is a boy (with short hair and a blue shirt) and a girl (with long hair, tying her hair back) present in the video. The distinction in hairstyle, clothing, and appearance supports this conclusion.

Sub-Q2: Can the person said to be the boy (in the blue shirt) be identified as male, and the person with long hair as female?

Sub-A2: The person in the blue shirt appears male and the one with long hair appears female based on visual cues, but these are assumptions based on stereotypes and not definitive without further context.

Sub-Q3: Is there any visible hair dryer present in any frame of the video?

Sub-A3: No hair dryer is visible in any frame of the video—only bottles and toiletries appear, confirming its absence.

Sub-Q4: Does the boy hold or handle an object that resembles a hair dryer at any point?

Sub-A4: The boy never handles an object resembling a hair dryer; instead, he uses something like a hairbrush on the girl's hair, confirming no hair dryer is present in the video.

Final Question: Is the boy using a hair dryer on the girl's hair?

Final Answer: The boy is not using a hair dryer on the girl's hair. No such device appears in any frame, and he uses only his hands or possibly a simple brush, confirming he's not drying her hair with a hair dryer. ✓

Question: Is the boy using a hair dryer on the girl's hair?

OpenAI GPT-4.1 Answer: Based on the frames, the boy is holding an object that visually resembles a hair dryer, and he is pointing it at the girl's hair in a manner consistent with using a hair dryer. Therefore, yes, the boy is using a hair dryer on the girl's hair. ✗

Question: Is the boy using a hair dryer on the girl's hair?

Gemini 2.5 Pro Answer: Yes, it appears the boy is using a hair dryer on the girl's hair. ✗ He is holding an object that resembles a hair dryer and pointing it at her hair, while she holds her hair up.

Figure 2. An example result from the CVRR-ES benchmark where Gemini 2.5 Pro and GPT-4.1 fail, but DIVE achieves the correct answer by breaking down the question and solving it iteratively.

reasoning, intent estimation, and video summarization using object detection, enabling highly accurate and contextually appropriate answers to complex questions in the CVRR-ES dataset. Experimental results show that our method achieves accuracies of 91.55% on the validation set and 81.44% on the test set, significantly outperforming existing methods and other participating teams. Furthermore, ablation studies confirmed the significant contribution of each module, such as Question Breakdown, Intent Estimation, and Video Summarization, to overall performance.

Future work will focus on further optimizing computational efficiency, exploring dynamic module selection, and evaluating the generalizability of our approach to other video QA benchmarks.

References

- [1] Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. Videoagent: A memory-augmented multimodal agent for video understanding. *ArXiv*, abs/2403.11481, 2024. 1
- [2] Muhammad Uzair khattak, Muhammad Ferjad Naeem, Jameel Hassan, Naseer Muzzamal, Federcio Tombari, Fahad Shahbaz Khan, and Salman Khan. How good is my video Imm? complex video reasoning and robustness evaluation suite for video-Imm. *arXiv:2405.03690*, 2024. 3, 4
- [3] Noriyuki Kugo, Tatsuya Ishibashi, Kosuke Ono, and Yuji Sato. Vdma: Video question answering with dynamically generated multi-agents. *ArXiv*, abs/2407.03610, 2024. 1
- [4] Noriyuki Kugo, Xiang Li, Zixin Li, Ashish Gupta, Arpandeeep Khatua, Nidhish Jain, Chaitanya Patel, Yuta Kyuragi, Yasunori Ishii, Masamoto Tanabiki, Kazuki Kozuka, and Ehsan Adeli. Videomultiagents: A multi-agent framework for video question answering. *ArXiv*, abs/2504.20091, 2025. 1
- [5] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C H Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *ICML*, 2023. 1
- [6] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. 2024.
- [7] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *Conference on Empirical Methods in Natural Language Processing*, 2023. 1
- [8] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer, 2024. 3
- [9] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*, 2024. 1
- [10] OpenAI. Introducing deep research. <https://openai.com/index/introducing-deep-research/>, 2025. Accessed: 2025-06-05. 1
- [11] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. *ArXiv*, abs/2312.02051, 2023. 1
- [12] Chuyi Shang, Amos You, Sanjay Subramanian, Trevor Darrell, and Roei Herzig. Traveler: A modular multi-Imm agent framework for video question-answering. In *Conference on Empirical Methods in Natural Language Processing*, 2024. 1
- [13] Chao-hong Wang et al. Videoagent: Long-form video understanding with large language model as agent. In *NeurIPS*, 2024. 1
- [14] Yi Wang, Kunchang Li, Yizhuo Li, Yanan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, Sen Xing, Guo Chen, Juntong Pan, Jiashuo Yu, Yali Wang, Limin Wang, and Yu Qiao. Internvideo: General video foundation models via generative and discriminative learning. *ArXiv*, abs/2212.03191, 2022. 1
- [15] Ying Wang, Yanlai Yang, and Mengye Ren. Lifelongmemory: Leveraging llms for answering queries in long-form egocentric videos. In *arXiv*, 2023. 1
- [16] Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. Videotree: Adaptive tree-based video representation for llm reasoning on long videos. *ArXiv*, abs/2405.19209, 2024. 1
- [17] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Zero-shot video question answering via frozen bidirectional language models. *ArXiv*, abs/2206.08155, 2022. 1
- [18] 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 *Conference on Empirical Methods in Natural Language Processing*, 2023. 1
- [19] Haoyu Zhang, Yuquan Xie, Yisen Feng, Zaijing Li, Meng Liu, and Liqiang Nie. Hcqa @ ego4d egoschema challenge 2024. *ArXiv*, abs/2406.15771, 2024. 1
- [20] Zhuo Zhi, Qiangqiang Wu, Minghe shen, Wenbo Li, Yinchuan Li, Kun Shao, and Kaiwen Zhou. Videoagent2: Enhancing the llm-based agent system for long-form video understanding by uncertainty-aware cot. *ArXiv*, abs/2504.04471, 2025. 1