In this Appendix, we will first detail the implementation of object re-ID method. Then, the tasks included in *VideoAgent* and the corresponding models will be listed, followed by the experiment settings of *VideoAgent* and all the comparative methods. Finally, cases of the inference of *VideoAgent* will be illustrated.

A Object Re-Identification

Based on the tracking results, object re-identification (re-ID) aims at merging the occurrences of the same object in different period (diverse tracking IDs). The following algorithm shows the procedure of object re-ID. It receives a set of tracking IDs, and output a set of Re-ID groups, where each Re-ID group contains several tracking IDs that belong to the same object, representing a unique object ID in the database.

Algorithm 3: Object Re-Identification by Grouping.			
Input: video V, tracking IDs $\{t_1, t_2,, t_n\}$			
Output: a list of RE-ID groups $G = \{U_1, U_2,, U_m\}$			
1 Initialize tracking IDs $T = \{t_1, t_2,, t_n\}$ to be examined;			
2 Initialize the set of re-ID groups $G = \{\};$			
3 for frame f in V do			
4 for t_i appears in f and $t_i \in T$ do			
5 for Re-ID group U in G do			
6 if $\forall t_j \in U$, share-no-frame (t_i, t_j) and $\forall t_j \in U$, $sim(t_i, t_j) > 0.5$			
and $\exists t_j \in U, sim(t_i, t_j) > 0.62$ then			
7 remove t_i from T ;			
8 add t_i to U ;			
9 break;			
10 If $t_i \in I$ then			
remove t_i from T ;			
create a new group $U = \{t_i\};$			
13 add U to G ;			
14 output G ;			

For each video frame, the algorithm checks every tracking ID in the frame that has not been examined and try to assign it to any existing Re-ID group. A tracking ID t_i should satisfy three conditions in order to be merged to a Re-ID group U: 1) It should not co-exist with any tracking IDs in U, since the same object only has one bounding box in each frame; 2) It should has $sim(t_i, t_j) > 0.5$ for all t_j in U; where sim refer to the CLIP and DINOv2 feature similarity in the paper; 3) At least one tracking ID t_j in group U satisfies $sim(o_i, o_j) > 0.62$. If the

tracking ID t_i cannot be merged to any existing Re-ID group, then the algorithm will spare a new re-ID group initialized with t_i . The results of object re-ID are used to construct the SQL database, with each re-ID group corresponding to a unique object ID in the database.

B Tasks and Models

Tab. 7 shows the different tasks in *VideoAgent* and their corresponding models. For each task, the granularity level of the details is also shown. For instance, in the task of segment captioning, the details of captions usually include the actions of the characters and the primary objects in the video that the characters are interacting with.

Table 7: The methods and the granularity-level of the extracted information in differenttasks.

Task	\mathbf{Method}	Detail Granularity	
Memory			
Segment Captioning	LaViLa	action, primary object	
Object Tracking	RT-DETR+ByteTrack	object category	
Object Re-ID	CLIP+DINOv2	object feature	
Tools			
Video Grounding	ViCLIP+Text-Embedding-3-Large	action, primary object	
Visual Question Answering	Video-LLaVA	action, object	



C Experiment Settings

C.1 Settings of VideoAgent

Prompt of *VideoAgent* The tool-use capabilities of the LLM (GPT-4) is facilitated using LangChain⁴. The LLM is prompted by the following text for the video question answering task.

You are tasked with answering a multiple-choice question related to a video. The question has 5 choices, labeled as 0, 1, 2, 3, 4. The video is segmented into 2-second segments, each with an integer ID starting from zero and incrementing in chronological order. Each segment has a caption depicting the event. There is an object memory that records the appearing objects in each segment. The object memory is maintained by another agent. You have access to the following tools:

 $\{tools\}$

ATTENTION:

1. the segment captions with prefix '#C' refer to the camera wearer, while captions with prefix '#O' refer to someone other than the camera wearer. 2. You can use both 'visual_question_answering' and 'object_memory_querying' to answer questions related to objects or people.

3. The 'visual_question_answering' may have hallucination. You should pay more attention to the description rather than the answer in 'visual question answering'.

4. The input to the tools should not contain the name of any other tool as well as the token '.

5. Its easier to answer the multiple-choice question by validating the choices.

6. If the information is too vague to provide an accurate answer, make your best guess.

Use the following format:

Question: the input question you must answer Thought: you should always think about what to do Action: the action to take, should be one of [{tool_names}] Action Input: the input to the action Observation: the result of the action... (this Thought/Action/Action Input/Observation can repeat N times) Thought: I now know the final answer

⁴ https://www.langchain.com/

Final Answer: the correct choice label (0, 1, 2, 3, 4) to the original input question

Begin!

Question: {input} Thought: {agent scratchpad}

In the above prompt format, **tools** refer to a set of tool names and their functional description, including:

caption_retrieval: Given an input tuple (start_segment, end_segment), get all the captions between the two segment IDs, 15 captions at most. end segment<start segment+15.

segment_localization: Given a single string description, this tool returns the total number of segments and the top-5 candidate segments with the highest caption-description similarities.

visual_question_answering: Given an input tuple (question, segment_id), this tool will focus on the video segment starting from segment_id-1 to segment_id+1. It returns the description of the video segment and the answer of the question based on the segment.

object_memory_querying: Given an object-related question such as 'what objects are in the video?', 'how many people are there in the video?', this tool will give the answer based on the object memory. This tool is not totally accurate.

input refers to the multiple-choice question input, including a question and 5 options. agent _scratchpad is a list maintained by LangChain that stores the intermediate steps of the agent.

Object Memory Querying The object_memory_querying tool is achieved by another LLM agent (GPT-4) specialized in SQL writing, equipped with the following tools:

- database_querying(*program*): return the results from the object memory database by executing the SQL *program*.
- open_vocabulary_retrieval(description): return the possible object IDs that satisfy the object description.

Given an object-related query raised by the central agent, the memory agent will get the relevant object IDs by open vocabulary retrieval, translate the query into SQL program, fetch the results from the database in object memory by running the SQL program, and return the natural language answer to the central agent.



Experiment Settings of *VideoAgent* For NExT-QA and EgoSchema, we use the above prompt for testing the performance of *VideoAgent*. For Ego4D NLQ, the ensemble proportion of video-text and text-text similarities for LaViLa+VICLIP is 18:11, and that for Ego4D+ViCLIP is 7:8. The ensemble proportions is found by grid search on the training set of Ego4D NlQ according to the maximal overall performance on R1@0.3, R1@0.5, R5@0.3 and R5@0.5.

C.2 Settings of Comparative Methods

In the experiments, we test the performance of the following methods by our own. The experiment settings for different comparative methods are detailed as follows.

- SeViLA: The default settings provided in their code are adopted for evaluation. The video frame number is set to 32, and the key frame number is set to 4.
- Video-LLaVA: The default settings provided in their code are adopted for evaluation. The input frame number is set to 8.
- mPLUG-Owl: We follow the evaluation procedure in EgoSchema dataset paper [17], which prompts mPLUG-Owl by 'Given question <question text>, is answer <answer text> correct?' along with the video frames. The option with the highest softmax score of the token 'Yes' in the output text will be viewed as the answer of mPLUG-Owl. The input frame number is set to 5 according to the best mPLUG-Owl settings provided in EgoSchema dataset paper.
- ViperGPT: GPT-3.5 is adopted as the code generator. 4 frames are uniformly sampled from the video and the generated code is run on the 4 frames to gather information for answering the question.

D Case Study

In this section, the successful cases of *VideoAgent* on both NExT-QA and EgoSchema are illustrated. In each step towards the final answer, the LLM first reasons about the action to take, and then outputs the action (tool) and its input. The tool will be executed and return the textual results to the LLM. This procedure will iterate until the LLM gets the final answer.

D.1 Case 1

In this case, the LLM uses the tool segment_localization, caption_retrieval and visual_question_answering to answer the question. Due to the man in red only occupies a tiny area of the screen, Video-LLaVA does not find the man and produce an answer with hallucination to the question "what does the man do next". However, the LLM will synthesize all the information and produce a correct answer. The video can be found here⁵.



Question: "what does the man in red do after looking at a distance from the plane at the start"

- 0: "shakes his head"
- 1: "happy"
- 2: "point forward"
- 3: "count down"
- 4: "moves away"

Entering new AgentExecutor chain...

I need to find the segment where the man in red is looking at a distance from the plane. Then I can use the visual_question_answering tool to find out what he does next.

Action: segment_localization

Action Input: "man in red looking at a distance from the plane" Results: There are 44 segments in total, ranging from 0 to 43. 22: '#C C stares the drone on the ground', 15: '#C C looks around the area', 41: '#C C looks around the area', 39: '#O A man x adjusts a drone on the', 40: '#C C stares the drone on the'

⁵ https://youtu.be/5tCWCmCWJKw?si=-IKmlA20_2SqaI_W

The segment with the highest similarity to the description is segment 22, but it doesn't mention a plane. The next highest is segment 15, but it also doesn't mention a plane. The segments 39 and 40 mention a drone, which could be the "plane" referred to in the question. I'll check the captions around these segments to get more context.

 ${\bf Action: \ caption_retrieval}$

Action Input: (37, 42)

Results: {37: '#C C looks around the parking', 38: '#C C looks around the', 39: '#O A man x adjusts a drone on the', 40: '#C C stares the drone on the', 41: '#C C looks around the area', 42: '#C C looks around the'}

The captions from segments 37 to 42 show a man adjusting a drone and then staring at it. This could be the "man in red looking at a distance from the plane" referred to in the question. I'll use the visual_question_answering tool to find out what he does next.

 ${\bf Action: visual_question_answering}$

Action Input: ("what does the man do next?", 40)

Results: Description: The video is about a large airplane that is parked on the runway. It shows the plane from different angles, including a close-up of the front of the plane. Answer: The man in the video walks towards the plane and opens the door.

The man in the video walks towards the plane and opens the door. This action is not listed in the choices. The closest option is "moves away", which could be interpreted as the man moving away from his original position towards the plane.

Final Answer: 4

D.2 Case 2

In this case, segment_localization and visual_question_answering are used to answer the questions. Although some details provided by Video-LLaVA are incorrect, the LLM is able to choose the correct answer based on the captions of the video and the results from Video-LLaVA. The video can be found here⁶.



Question: "how does the boy keep the bat stable on the ground" 0: "lean against his body"

- 1: "sit on cushion"
- 1: "sit on cusnion
- 2: "put on stone"
- 3: "swing chair is protected"
- 4: "wear shoes"

Entering new AgentExecutor chain...

The question is asking about a specific action performed by a boy in the video. I need to find the segment where the boy is interacting with the bat and then use the visual_question_answering tool to understand how he keeps the bat stable. First, I will use the segment_localization tool to find the segments where the boy and the bat are mentioned together. Action: segment_localization

Action Input: "boy bat"

Results: There are 12 segments in total, ranging from 0 to 11. {7: '#O A man Y holds the baseball bat with both', 6: '#O The man B hits the ball with the rake in his right hand', 9: '#O man Y holds a head warmer', 10: '#O A man Z wears a helmet on his head with both hands', 2: '#O A

⁶ https://youtu.be/vzlPCFqdtQQ?si=X9vATb1ClBVM8oMM

man X walks around the field'}

The segment_localization tool has returned several segments, but only segment 7 seems to be relevant to the question as it mentions a man (possibly the boy) holding a baseball bat. I will use the visual_question_answering tool on this segment to get more details about how the boy keeps the bat stable.

Action: visual_question_answering

Action Input: ("how does the boy keep the bat stable?", 7)

Results: Description: The video is about a man who is practicing his archery skills in a park. He is seen holding a bow and arrow and shooting at a target. Answer: The boy keeps the bat stable by holding it firmly in his hand and keeping it steady. He also uses his body movements to keep the bat stable.

The visual_question_answering tool provided an answer that the boy keeps the bat stable by holding it firmly in his hand and using his body movements. This corresponds to the first option in the multiple-choice question.

Final Answer: 0

D.3 Case 3

In this case, the visual_question_answering tool contributes most to the successful question answering, since it identifies the wedding ring on the woman's left hand, although it mistakes the woman's action in the video (delivering a talk to others) as 'playing rock paper scissors with a man'. The video can be found here⁷.



Question: "how do you know that the woman with black spectacles at the start is married"

- 0: "wearing a wedding gown"
- 1: "holding flowers"
- 2: "wearing a wedding ring"
- 3: "grandparent and grandchild"
- 4: "walking with the groom"

Entering new AgentExecutor chain...

I need to find the segment where the woman with black spectacles appears at the start of the video. Then I can use the visual_question_answering tool to determine if she is married based on the given choices. Action: segment localization

Action Input: "woman with black spectacles at the start"

Results: There are 44 segments in total, ranging from 0 to 43. {30: '#O the woman A touches the glasses on her face with her left hand.', 11: '#O The woman X touches her glasses with her right hand', 14: '#O person X adjusts her glasses ', 16: '#O woman X looks at the television', 39: '#C C stares at a woman'}

⁷ https://youtu.be/EFMdUrOELqw?si=6iXTKJCW0x96a7Ap

The woman with black spectacles appears in the first segment (0). Now I will use the visual question answering tool to determine if she is married based on the given choices.

Action: visual question answering

Action Input: ("how do you know that the woman is married", 0)

Results: Description: The video shows a man and a woman playing a game of rock paper scissors. The woman wins, and the man gets angry. Answer: The video does not provide any information about the woman's marital status. However, it is possible to observe that she is wearing a wedding ring on her left hand.

The woman is wearing a wedding ring which indicates that she is married. This matches with choice 2. Final Answer: 2

D.4 Case 4

In this case, the object memory is queried by the LLM (achieved by a memory agent mentioned in Appendix C.1) to answer the quantity of the elephants, and output the correct answer. The video can be found here⁸.



Question: "how many elephants are there"

- 0: "one"
- 1: "four"
- 2: "three"
- 3: "six"
- 4: "two"

Entering new AgentExecutor chain..

I should use the 'object_memory_querying' tool to find out how many elephants are in the video. Action: object_memory_querying

Action Input: 'how many elephants are there in the video?' **Results**: There are 2 elephants in the video.

I now know the final answer. **Final Answer**: 4

⁸ https://youtu.be/OB87EvVi-as?si=DST7FJfGtPmW6ZcB



D.5 Case 5

This is a typical question in EgoSchema, which requires strong textual understanding and reasoning abilities of the model to correctly answer it. By retrieving the captions, the LLM can gather the information of the ingredients in various segments, and uses visual_question_answering' tool on the corresponding segment to investigate the function of each ingredient. The video can be found here⁹.



Question: 'Analyze the choice of ingredients and their importance in the video. how do they contribute to the final dish?',

0: 'The tissue paper is an important ingredient in the dish. it is used to clean up any spills.'

1: "The can cover, a crucial component in the dish, serves to maintain the minced beef's freshness effectively by protecting it."

2: 'The wax paper, being a crucial ingredient in the dish, is utilized effectively to securely wrap the can to prevent any potential leakage.'

3: 'The minced beef, spring onions, and mayonnaise are all important ingredients in the dish. the minced beef provides protein, the spring onions provide vitamins and minerals, and the mayonnaise provides flavor.'

4: "The transparent glass bowl serves as an important crucial ingredient in the dish's preparation. it is efficiently used to thoroughly mix the various ingredients together well.

Entering new AgentExecutor chain...

⁹ https://youtu.be/s0LU2K7Ukwo?si=RNEIfZjGCmWeM-1t

The question is asking about the importance of ingredients in the final dish. I need to find out what ingredients are used in the video and how they contribute to the final dish. I will start by using the caption_retrieval tool to get an overview of the events in the video.

Action: caption_retrieval

Action Input: (0, 14)

Results: $\{0: \ ' \# C \ C \ scoops \ sauce from the container with the spoon in$ her right hand', 1: '#C C pours the tomato paste from the spoon into the bowl of macaroni and cheese with her right hand', 2: '#C C scoops sauce from the sauce container with the spoon in her right hand', 3: #CC pours the spice into the plate of scrambled eggs with the spoon in her right hand', 4: '#C C applies the ketchup from the spoon in his right hand on the salad in the bowl', 5: #C C scoops sauce from the sauce container with the spoon in her right hand', 6: '#C C pours the tomato sauce into the bowl on the dining table with the spoon in her right hand', 7: #C C pours milk from the bottle in her left hand into the cup of milk in her right hand', 8: '#C C puts some sauce from the sauce container on the spoon with her right hand.', 9: '#C C holds the spoon in the sauce container with her left hand.', 10: '#C C drops the container on the table with her left hand', 11: '#C C closes the trash', 12: '#C C walks towards the dining table', 13: '#C C drops the tissue paper on the table with her left hand', 14: '#C C picks a tissue paper from the floor with her left'}

The captions indicate that the dish involves sauce, tomato paste, macaroni and cheese, spice, scrambled eggs, ketchup, salad, tomato sauce, milk, and tissue paper. The tissue paper seems to be used for cleaning rather than as an ingredient. I need to find out more about how these ingredients are used. I will use the visual_question_answering tool to get more details about the use of these ingredients.

Action: visual_question_answering

Action Input: ('What is the role of the sauce in the dish?', 0)

Results: Description: The video is about a person preparing a meal in a kitchen. They are seen cutting up food and mixing it in a bowl. Answer: The sauce is poured over the noodles and mixed in with the other ingredients, adding flavor and texture to the dish.

The sauce is used to add flavor and texture to the dish. Now, I will use the visual_question_answering tool again to find out more about the role of the other ingredients.

Action: visual_question_answering

Action Input: ('What is the role of the spice and scrambled eggs in the dish?', 3)

Results: Description: The video is about a person preparing and eating a meal in a spacious and well-lit room. Answer: The spice and scrambled

eggs are added to the dish to enhance the flavor and texture. The spice adds a burst of flavor and aroma, while the scrambled eggs provide a creamy and tender texture that complements the other ingredients.

The spice and scrambled eggs are used to enhance the flavor and texture of the dish. Now, I will use the visual_question_answering tool again to find out more about the role of the remaining ingredients.

Action: visual_question_answering

Action Input: ('What is the role of the tomato sauce and milk in the dish?', 6)

Results: Description: The video is about a person preparing and eating a meal in a spacious and well-lit kitchen. The person is seen cutting up food on a table, mixing it in a bowl, and then eating the food. Answer: The tomato sauce and milk are poured over the ice cream, adding a sweet and tangy flavor to the dish.

The tomato sauce and milk are used to add a sweet and tangy flavor to the dish. However, the answer seems to be hallucinating about ice cream, which was not mentioned in the captions. I will use the visual_question_answering tool again to find out more about the role of the remaining ingredients.

Action: visual question answering

Action Input: ('What is the role of the tomato paste and macaroni and cheese in the dish?', 1)

Results: Description: The video is about a person preparing and eating a meal in a spacious and well-lit room. Answer: The tomato paste and macaroni and cheese are added to the dish to enhance its flavor and texture. The tomato paste adds a rich and savory taste, while the macaroni and cheese provides a creamy and comforting element.

The tomato paste and macaroni and cheese are used to enhance the flavor and texture of the dish. The tomato paste adds a rich and savory taste, while the macaroni and cheese provides a creamy and comforting element. Now, I will use the visual_question_answering tool again to find out more about the role of the tissue paper.

Action: visual_question_answering

Action Input: ('What is the role of the tissue paper in the dish?', 13) Results: Description: The video is about a person cleaning a bathroom. Answer: The tissue paper is used to clean the dish.

The tissue paper is used for cleaning, not as an ingredient in the dish. Now, I can compare the information I have gathered with the choices provided in the question.

Final Answer: 3