



Embodied VideoAgent: Persistent Memory from Egocentric Videos and Embodied Sensors Enables Dynamic Scene Understanding

Yue Fan^{1*}, Xiaojian Ma^{1*†}, Rongpeng Su^{1,2}, Jun Guo^{1,3}, Rujie Wu^{1,4}, Xi Chen¹, Qing Li^{1†}

¹State Key Laboratory of General Artificial Intelligence, BIGAI, Beijing, China

²University of Science and Technology of China ³Tsinghua University ⁴Peking University

{fanyue,maxiaojian,liqing}@bigai.ai

<https://embodied-videoagent.github.io>

Abstract

This paper investigates the problem of understanding dynamic 3D scenes from egocentric observations, a key challenge in robotics and embodied AI. Unlike prior studies that explored this as long-form video understanding and utilized egocentric video only, we instead propose an LLM-based agent, Embodied VideoAgent, which constructs scene memory from both egocentric video and embodied sensory inputs (e.g. depth and pose sensing). We further introduce a VLM-based approach to automatically update the memory when actions or activities over objects are perceived. Embodied VideoAgent attains significant advantages over counterparts in challenging reasoning and planning tasks in 3D scenes, achieving gains of 4.9% on Ego4D-VQ3D, 5.8% on OpenEQA, and 11.7% on EnvQA. We have also demonstrated its potential in various embodied AI tasks including generating embodied interactions and perception for robot manipulation. The code and demo will be made public.

1. Introduction

Understanding dynamic 3D scenes is crucial to the development of generally capable embodied AI [17, 51–53]. In this paper, we investigate approaching this problem using egocentric observations [5, 7, 11, 20, 26, 33, 37], which is one of the most intuitive way of how humans and robots perceive the world around them. The key challenges include: 1) Making sense of environments from lengthy egocentric videos and other forms of embodied sensory inputs (depth maps, camera poses, etc.) [17, 41, 44]; 2) Handling dynamic environments as actions and activities might be performed by embodied agents themselves and other co-

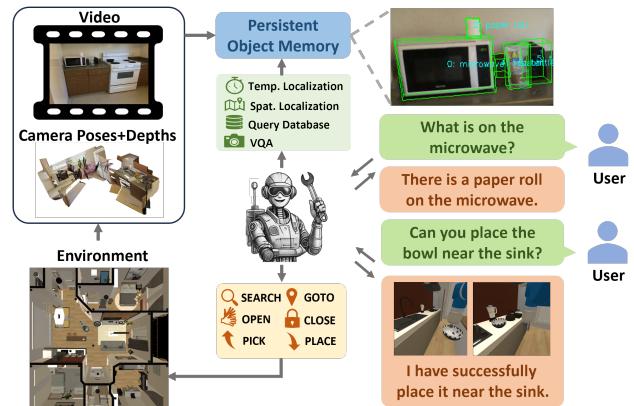


Figure 1. *Embodied VideoAgent* is a multimodal agent that 1) builds scene memory from both egocentric video and embodied sensory input; 2) utilizes multiple tools to query this memory; 3) activates embodied action primitives to interact with the environments, effectively fulfills various user requests.

habited characters [5, 11, 33]; 3) Maintaining a persistent memory about the scene that allows frequent update over time [12, 18, 40]. However, existing efforts on this front mostly adopt end-to-end pretrained multimodal large models (MLMs) [17, 24, 25, 27, 62, 64, 65]. Their capabilities of handling long-form videos and embodied sensory observations have been questioned by several prior studies [23, 34, 46, 56], especially when the scene being depicted is highly volatile with complex events and spatial-temporal dependencies [15, 19, 20]. Some sophisticated MLMs coming out recently have attained great progress in understanding long-form videos and the underlying embodied scenes [47, 48, 54], but the computation cost can grow at a prohibitively expensive rate. All these issues have hindered the progress of deploying robust dynamic scene perception pipelines on edge devices like robots in the wild.

Unlike these end-to-end models, there has been rapid development in another family of multimodal understanding

*Equal contribution.

†Corresponding authors.

approaches – *multimodal agents* [7, 9, 14, 45, 55]. These methods utilize the advanced reasoning and *tool-usage* ability of pretrained large language models (LLMs) and solve intricate multimodal tasks by calling several tool models (object understanding, question answering, etc.) interactively, alleviating the issue of expensive training and inference cost. Recently, they have been ported to long-form video understanding and have demonstrated remarkable performances and cost-efficiency over end-to-end counterparts [7, 9, 45]. Their key idea is to construct a temporal memory from the video and invoke several tools to query the memory. However, extending them to understanding dynamic 3D scenes is **non-trivial**. The challenges as mentioned earlier require the system to have a comprehensive yet precise understanding of objects in the scene subjected to constant change due to various actions and activities being performed by the embodied agents and other characters. Our early explorations (see [Section 3](#)) on simply applying these systems (*e.g.* VideoAgent [7]) to dynamic 3D scenes in embodied environments have suggested that merely constructing scene memory from video with hand-crafted pipelines cannot meet the aforementioned requirements, especially on the precise understanding of objects and the support of dynamic memory update, leading to unsatisfactory performances on these tasks.

To this end, we propose *Embodied VideoAgent*, a simple yet effective multimodal agent for understanding dynamic 3D scenes. Our agent is based upon VideoAgent [7], a recent multimodal agent that can solve various video understanding tasks by constructing memory on long-form videos and performing LLM-based queries over the memory. Our **key innovation** is to augment it with two novel designs for dynamic scenes in embodied environments: 1) a **persistent object memory** that is constructed from *both* egocentric video and embodied sensory input (depth maps and camera poses). Fusing video with these modalities could help build more precise memory on scene objects, which is crucial to embodied reasoning and planning; 2) a **VLM-based memory update mechanism** that automatically identifies relevant objects and their state changes when actions and activities are being perceived, then promptly updates the corresponding entries in the persistent object memory. In addition to understanding tasks, we explore the potential of *Embodied VideoAgent* in generating embodied user-assistant interactions. Specifically, we introduce an LLM-based multi-agent framework [10], where a user agent proposes tasks, and an assistant agent (effectively an *Embodied VideoAgent*) progressively explores the scene to complete these tasks. The assistant provides feedback to the user while tracking its evolving understanding of the dynamic scene using persistent object memory.

We conduct extensive evaluations of *Embodied VideoAgent* on several embodied scene understanding tasks, including object localization from free-form queries in egocen-

tric views of dynamic scenes with Ego4D-VQ3D [11, 31], general question answering on embodied 3D scenes with OpenEQA [32], and question answering on long-form embodied robot-environment interactions with EnvQA [8]. We compare *Embodied VideoAgent* against both the canonical end-to-end multimodal LLMs and other multimodal agents. Results confirm the advantages of *Embodied VideoAgent*: achieving gains of 4.9% on Ego4D-VQ3D, 5.8% on OpenEQA, and 11.7% on EnvQA. Our further exploration has covered its applications in generating embodied interaction and perception for robot manipulation.

Our contributions can be summarized as follows:

- We propose a persistent object memory along with a VLM-based automatic memory update method to construct and maintain comprehensive yet precise memory of dynamic 3D scenes from both egocentric and embodied (depth maps, camera poses) sensory observations.
- With the resulting agent, *Embodied VideoAgent*, we further develop an LLM-based multi-agent framework that can produce embodied user-assistant interactions, where the user proposes tasks and the assistant (an *Embodied VideoAgent*) progressively explores the scene to complete these tasks.
- We conduct thorough evaluations of *Embodied VideoAgent* on various embodied scene understanding tasks against both end-to-end multimodal LLMs and multimodal agent baselines, siding with applications in two embodied AI tasks, demonstrating the effectiveness of *Embodied VideoAgent*.

2. Embodied VideoAgent

We illustrate the proposed *Embodied VideoAgent* in [Figure 2](#). Since our agent is based upon VideoAgent [7], we will first quickly recap its key ideas ([Section 2.1](#)), then move on to cover the key new memory and tool design, including persistent object memory and VLM-based memory update method ([Section 2.2](#)). Finally, we will detail how we utilize *Embodied VideoAgent* for generating embodied user-assistant interactions ([Section 2.3](#)).

2.1. Recap: VideoAgent

VideoAgent [7] adopts the following pipeline: given a video V sliced into n segments $[v_1, \dots, v_n]$, it first constructs a temporal memory \mathcal{M}_T , which captures the textual descriptions (and features) of each segment; and an object memory \mathcal{M}_O , which tracks and stores the occurrences of objects and persons in the video. Then for any incoming task, an LLM decomposes it into several subtasks and invokes tool models to query the temporal and object memory. Finally, the responses of all tool-calling will be aggregated and sent to an LLM for a final answer.

Temporal and Object Memory. For temporal memory, it is effectively a table with n rows, where n is the number

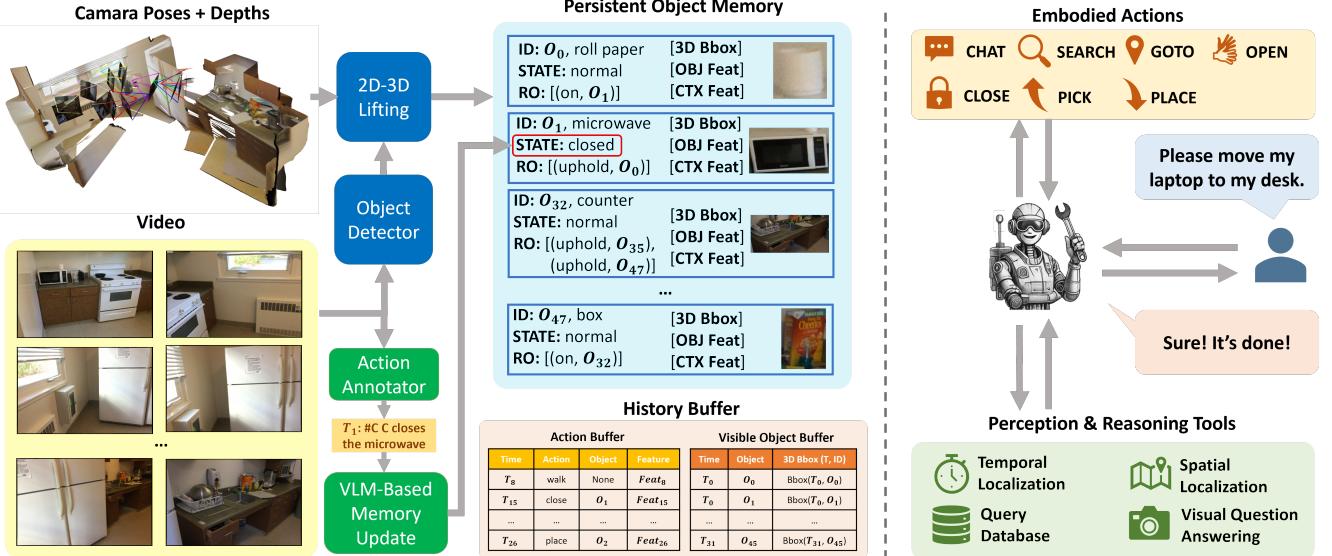


Figure 2. An overview of *Embodied VideoAgent*. Left: We first translate the egocentric video and embodied sensory input (depth maps and camera poses) into structured representations: *persistent object memory* and history buffer. While the memory can be updated using VLM to support dynamic scenes where actions are being performed constantly; Right: the LLM within *Embodied VideoAgent* is prompted to fulfill the user’s request by interactively invoking tools to query the memory and calling embodied action primitives to complete the task.

of short (2s) video segments. Each row has four columns: segment ID, caption of this segment s_{caption} , visual feature of this segment e_{video} and text embedding of the caption e_{caption} . For object memory, it includes a SQL database and a feature table of all identified objects in the video. The SQL database has three columns: unique object ID, object category, and the segment IDs where the object occurs. Its construction requires object detection, tracking, and re-ID. The feature table stores the CLIP feature of the object image.

Tool-usage and Inference. VideoAgent utilizes four tools – caption retrieval, segment localization, visual question answering, and object memory query to access the temporal and object memory. The inference process is straightforward: given an input query, VideoAgent selects an appropriate tool, invokes it, and stores the result in a buffer. This loop continues until VideoAgent either decides to stop or reaches a predefined maximum number of steps, after which it generates a final response based on the buffer’s content.

Readers are encouraged to refer to [7] for more details.

2.2. Memory and Tools of Embodied VideoAgent

Embodied VideoAgent adopts the following memory and tool design upon its predecessor VideoAgent: given an egocentric video downsampled to n frames $V = [I_1, \dots, I_n]$, with depth map and camera 6D pose of each frame $D = [(d_1, p_1), \dots, (d_n, p_n)]$ (d and p denote depth map and camera pose, respectively), it constructs the original temporal memory \mathcal{M}_T of VideoAgent (not shown in Figure 2), the newly introduced persistent object memory \mathcal{M}_O , and two simple history buffers. Four tools (query_db, temporal_loc, spatial_loc, vqa) can be in-

voked to access the memory. Several embodied action primitives are available to be called to interact with the physical environment. Details can be found below:

Persistent Object Memory \mathcal{M}_O . It maintains an entry for each perceived object in the 3D scene. Each object entry includes the following fields: a unique object identifier O_i with object category (**ID**), a state description of the object (**STATE**), a list of related objects and their relations (**RO**), 3D bounding box of the object(**3D Bbox**), visual feature of the object (**OBJ Feat**) and visual feature of the environment context where the object locates(**CTX Feat**). These fields provide comprehensive details of scene objects and their surroundings.

Construction of \mathcal{M}_O . Given an incoming 2D egocentric frame I_i , depth map d_i and camera pose p_i , we first use an open-vocabulary object detection model called YOLO-world [4] to extract objects and their categories (the **ID** field) from 2D frame I_i . Its state description will be initialized as “normal” (the **STATE** field). The CLIP feature of frame I_i and the cropped object picture using the 2D bounding box will become the **CTX Feat** field and the **OBJ Feat** field, respectively. Further, by utilizing the depth map d_i along with camera pose p_i , we can obtain the object’s 3D bounding box using 2D-3D lifting (projection) [13] (the **3D Bbox** field). Then we follow the prior practice [17, 21, 64] to extract the relations among the detected objects using their 3D bounding boxes (the **RO** field). So far, only two pairs of relations “on/uphold” and “in/contain” are considered. To avoid duplicated object entries in \mathcal{M}_O , an object re-ID procedure [7] is also conducted before inserting an object as a new entry. Objects are considered identical based on

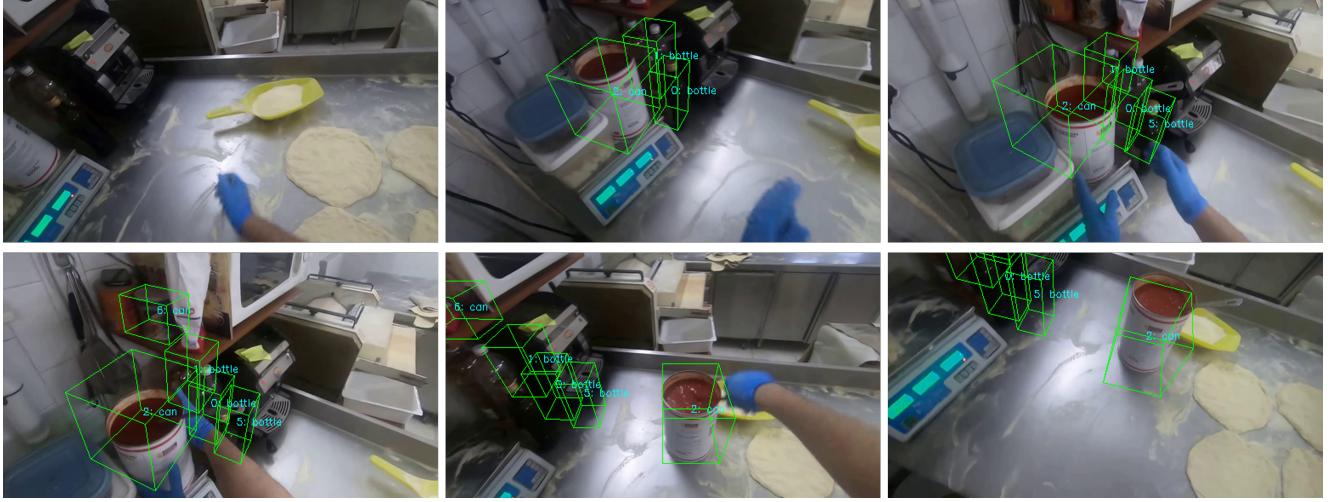


Figure 3. Visualization of the entries in *persistent object memory* \mathcal{M}_O . Each 3D bounding box corresponds to an entry in the memory. As the video proceeds, objects (e.g. the large canned tomato paste) can be tracked/re-IDed and have their memory entries updated.

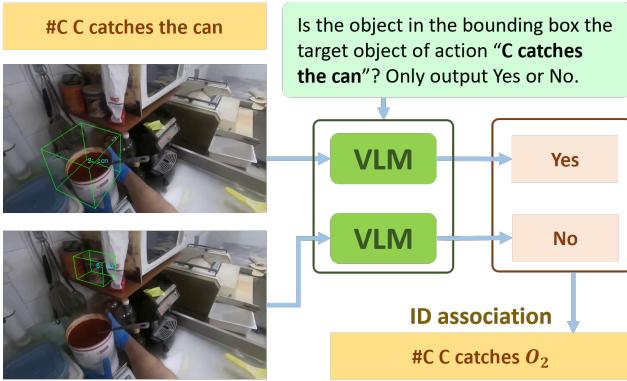


Figure 4. An illustration of our VLM-based memory update method. This approach effectively prompts the VLM to associate an action with relevant object entries in memory through visual prompting, identifying the entries corresponding to the action’s target objects.

their proximity in both visual appearance and 3D location. Once an object is re-IDed to an existing object entry, we will update its **3D Bbox**, **Obj Feat**, and **CTX Feat** fields using moving average, while the **RO** field will be re-computed along with other objects detected in the current frame. Due to space constraints, details of our re-ID algorithm and object entry update after re-ID are provided in *Appendix*. A visualization of how object entries are created and updated in \mathcal{M}_O can be found in [Figure 3](#).

Memory Update with VLM. A key challenge in persistent object memory lies in updating memory when actions are performed on objects, especially under conditions of visual occlusion (e.g., hand-can interaction in [Figure 3](#)). We address this issue by leveraging action information and vision-language models (VLMs). As shown in [Figure 4](#), when an action occurs (e.g., “C catches the can”), we first retrieve relevant object entries in \mathcal{M}_O associated with the “can” that are visible in the current frame (in this example, two entries). For each entry, we render its 3D bounding box onto

the frame and prompt the VLM to determine if the object within the box is the action’s target. Such visual prompting [1, 58] associates the action with corresponding entries in the object memory. Finally, we programmatically update these entries, such as modifying the **STATE** field to “in-hand” since the action is “catches the can”. Additional details on the programmatic update are provided in *Appendix*.

History Buffer. In addition to persistent object memory that provides real-time information on current scene objects, we found that maintaining a simple record of past perception and action history further enhances dynamic scene understanding. For this purpose, we introduce two history buffers: an **action buffer**, which logs each action performed along with the action timestamp, action name, target object ID (identified using the VLM-based method), and the CLIP feature of the current frame; and a **visible object buffer**, which logs each detected object along with the detection timestamp, object ID, and 3D bounding box. These buffers are also referenced by the tools described later.

Tools and Embodied Action Primitives. We equip *Embodied VideoAgent* with four tools: `query_db()`, which processes natural language queries to retrieve the top-10 matching object entries by searching both the persistent object memory and history buffers; `temporal_loc()`, inherited from *VideoAgent*, which maps natural language queries to specific video timesteps; `spatial_loc()`, which provides a 3D scene location (aligned with the camera’s coordinate system) based on object and room queries; and `vqa()`, which answers open-ended questions about a given frame. Additionally, the agent can perform seven embodied action primitives: `chat()` for user interaction; `search()` to conduct exhaustive scene searches for specified objects; `goto()` for location navigation; and `open()`, `close()`, `pick()`, and `place()` for object interactions. Further implementation details on tools and action primitives

are provided in *Appendix*.

Note on camera poses. While readers may view the requirement for precise 6D camera poses as idealized for real-world embodied agent settings, *Embodied VideoAgent* demonstrates robustness to pose estimation noise. In our experiments, camera poses for Ego4D-VQ3D ([Section 3.1](#)) and EnvQA ([Section 3.2.3](#)) are estimated using COLMAP [43] and DUST3R [49], respectively, which are inherently noisier than the ground truth poses available in OpenEQA ([Section 3.2.2](#)). Despite these variances, our agent consistently achieves substantial improvements over baselines (e.g., VideoAgent) across all three settings. In *Appendix*, we also present additional results on OpenEQA using estimated noisy poses to further substantiate this robustness. We hypothesize that our memory and tool design provide redundancy, enabling tasks to be completed via multiple pathways, effectively bypassing potentially flawed memory entries or tools. More inference examples are provided in *Appendix*.

2.3. A Two-Agent Framework for Generating Embodied Interactions

Collecting synthetic data for training foundation models, particularly embodied foundation models, has recently gained considerable interest [28]. We explore a novel approach with *Embodied VideoAgent* to gather synthetic embodied user-assistant interaction data. This dataset comprises episodes where a user interacts with an assistant within embodied environments. Drawing inspiration from prior multi-LLM-agent research [10], we use one LLM to emulate the user’s role, while *Embodied VideoAgent* assumes the assistant’s role, exploring the environment and fulfilling the user’s diverse requests. An overview of this framework is shown in [Figure 5](#). The user is prompted to propose varied and engaging tasks based on its limited scene graph knowledge—achieved by randomly trimming the full scene graph to stimulate curiosity—and the assistant’s feedback. Detailed prompting strategies are provided in *Appendix*.

3. Capabilities and Analysis

We evaluate *Embodied VideoAgent* on various dynamic and embodied scene understanding tasks, including 3D object localization in dynamic scenes using Ego4D-VQ3D ([Section 3.1](#)), embodied question answering with OpenEQA ([Section 3.2.2](#)), and general question answering over embodied interactions on EnvQA ([Section 3.2.3](#)). The performances are compared against state-of-the-art multimodal LLMs and multimodal agents. In [Section 3.3](#), we demonstrate its application to two embodied AI tasks: generating embodied interactions and perception for robot manipulation.

3.1. 3D Object Localization

We test *Embodied VideoAgent* on Ego4D Visual Queries 3D localization (VQ3D)[11]. Given an egocentric video

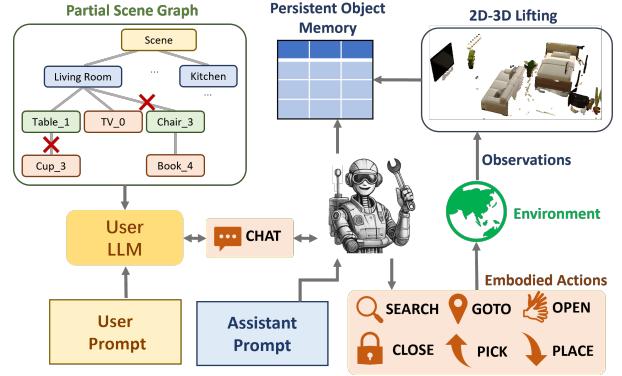


Figure 5. An overview of our synthetic embodied data collection framework. An LLM plays the user role and is prompted to propose engaging tasks based on a partial scene graph and the user’s feedback, while the user, effectively a *Embodied VideoAgent*, explores the scene and fulfills the user’s requests.

depicting how a human subject interacts with a dynamic environment, an image of a target object, and a query frame, the task of VQ3D is to output the position of the target object at the time stamp of the query frame.

3.1.1. Settings

Baselines. Two types of *Embodied VideoAgent* are tested: 1) retrieving the object with the highest visual score with the target image from the up-to-date object memory, denoted as *Embodied VideoAgent* (image); 2) retrieving the object with the same category as the target object, referred as *Embodied VideoAgent* (text). EgoLoc[31], Ego4D*[31], *Embodied VideoAgent* (text), and *Embodied VideoAgent* (image) are all based on the same precomputed camera poses and depth images provided by EgoLoc[31], the 1st place on VQ3D challenge. Ego4D[11] denotes the baseline method in the benchmark paper.

Metrics. Among the metrics, Succ% is the most important one that evaluates the success rate on all queries. Succ* (success rate on the answered queries only) and L2 (the average distance error) are all computed on the queries where the target object is detected by the method. The proportion of the answered queries to all queries is denoted as QwP%.

3.1.2. Results on VQ3D

[Table 1](#) shows the results on VQ3D validation set. Overall, *Embodied VideoAgent* (image) achieves the highest success rate, surpassing EgoLoc by 5%. We made the following observations:

Open-vocabulary object detector provides more candidate objects.

The higher QwP% rate of *Embodied VideoAgent* (image): 92.07% compared to that of EgoLoc, indicating the strong and robust performance of open-vocabulary object detection of *Embodied VideoAgent* empowered by YoloWorld[4]. The better Succ*% and L2 of EgoLoc can be attributed to its high-confidence predictions since these two metrics only evaluate the predicted queries. By contrast,

Table 1. Results of 3D object localization within dynamic scenes on the validation set of Ego4D-VQ3D[11].

Ego4D VQ3D				
Method	Succ% \uparrow	Succ*% \uparrow	L2 \downarrow	QwP% \uparrow
EgoLoc	80.49	98.14	1.45	82.32
Ego4D*	73.78	91.45	2.05	80.49
Ego4D	1.22	30.77	5.98	1.83
E-VideoAgent(text)	53.05	94.57	2.00	56.10
E-VideoAgent(image)	85.37	92.72	1.86	92.07

Embodied VideoAgent (image) sacrifices little Succ*% and L2 for more aggressive predictions on hard open-vocabulary queries, which finally results in the best QwP% and Succ%. **Visual Similarity is crucial for Object re-ID in a dynamic scene.** By only considering text for object retrieval, *Embodied VideoAgent* (text) has decent performance on VQ3D compared to Ego4D baseline, though not being competitive to *Embodied VideoAgent* (image). This can be attributed to the in-door settings of Ego4D videos, where functional objects (scissors, screwdrivers, etc) are usually clustered within the distance error of a successful detection. Applying visual-based object re-ID on object candidates boosts the performance, indicated by the large margin between *Embodied VideoAgent* (image) and *Embodied VideoAgent* (text), illustrating the effectiveness of the visual similarity score for object re-ID in dynamic scenes.

3.2. Embodied Question Answering

Given an embodied episode in a scene, Embodied Question Answering requires the model to answer the question about the scene and embodied activities, such as “what is the orange thing on the shelf to the right”, “where did I leave my remote controller”, etc. *Embodied VideoAgent* is tested on OpenEQA[32] and EnvQA [8], two recent benchmarks on open-ended embodied question answering.

3.2.1. Settings

Baselines. We equip *Embodied VideoAgent* with the four perception tools mentioned in Section 2.2. For vqa tool, we tested InternVL2-8B[2, 3] and GPT-4o, denoted as *Embodied VideoAgent* (InternVL2-8B) and *Embodied VideoAgent* (GPT-4o) respectively. Please note we compare with **zero-shot** baselines only following prior practices. On OpenEQA[32], the baseline methods include 1) Large Video Language Models: Video-LLaVA[25] and LLaMA-VID[24]; 2) multi-modal Agents: VideoAgent[7], GPT-4 w/LLaVA-1.5[32] (which leverages frame captions) and GPT-4 w/CG[32] (which uses scene graph information). *Embodied VideoAgent* is tested on a subset of the original dataset due to cost issues, with the subset size being one-fifth of the original dataset. The questions in the subset are randomly selected. On EnvQA[8], we tested *Embodied VideoAgent* with Video-LLaVA[25], LLaMA-VID[24] and VideoAgent[7]. We tested these methods on three types of questions of En-

Table 2. Results of embodied question answering on the EM-EQA split of OpenEQA[32]. Some scores are borrowed from the original benchmark paper.

OpenEQA			
Method	ScanNet	HM3D	ALL
GPT-4 w/ LLaVA-1.5	45.4	40.0	43.6
GPT-4 w/ CG	37.8	34.0	36.5
Video-LLaVA	41.5	34.6	39.2
LLaMA-VID	33.4	34.0	33.6
OpenEQA Subset			
Method	ScanNet	HM3D	ALL
Video-LLaVA	32.9	27.8	30.6
LLaMA-VID	31.2	28.0	29.4
VideoAgent	37.6	34.6	36.3
E-VideoAgent(InternVL2-8B)	39.7	43.0	41.2
E-VideoAgent(GPT-4o)	46.0	48.2	47.0

Table 3. Results of open-ended question answering over embodied interactions on the test set of EnvQA [8].

EnvQA			
Method	Events	Orders	States
Video-LLaVA	10.19	39.00	18.50
LLaMA-VID	9.98	54.00	5.50
VideoAgent	5.54	65.5	12.5
Embodied VideoAgent	25.91	68.00	35.50

vQA: States (e.g. “Where was the book moved?”), Events (e.g. “what happened, after throwing soap bar and before throwing soap bar to hit shower door?”), and Orders (e.g. “filling pot with water or use up soap bottle, which happened first”), with each type containing 200 questions.

3.2.2. Results on OpenEQA

Table 2 shows the results. It can be inferred that OpenEQA Subset is harder than the full OpenEQA validation set from the performance drops of Video-LLaVA and LLaMA-VID. On the hard subset, the two variants of *Embodied VideoAgent* both achieve good performances. Specifically, *Embodied VideoAgent* (GPT-4o) obtained 46.0% on ScanNet and 48.2% on HM3D, surpassing Video-LLaVA by 13.1% and 20.4% on the ScanNet and HM3D respectively.

Temporal localization + vqa tool solves embodied questions better than scene graph. The performance gaps between GPT4-w/LLaVA-1.5 and GPT-4 w/CG on the full OpenEQA indicate that LLM can better utilize frame captions for question answering other than scene graphs, which is validated by the better performance of *Embodied VideoAgent* (+16.4% over Video-LLaVA on subset) over GPT-4 w/CG (+4.4% over Video-LLaVA on full set). *Embodied VideoAgent* does not explicitly construct a complex scene graph during memory construction. Instead, for re-

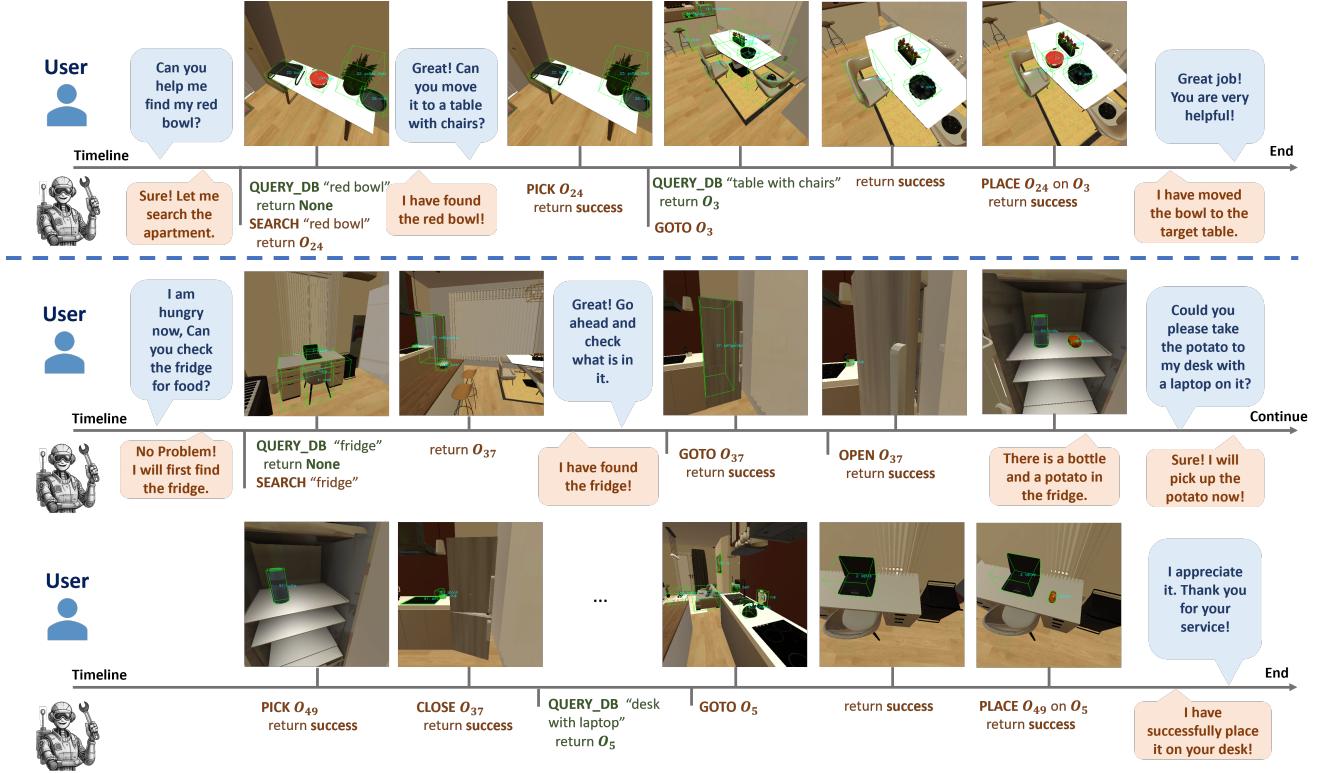


Figure 6. An episode of generated embodied user-assistant interaction. The episode is produced by the framework mentioned in Section 2.3, where an LLM plays the user and *Embodied VideoAgent* is the assistant. The episode comprises various embodied problem-solving that requires precise memory of the scene objects and tool usage. More example episodes can be found in Appendix.

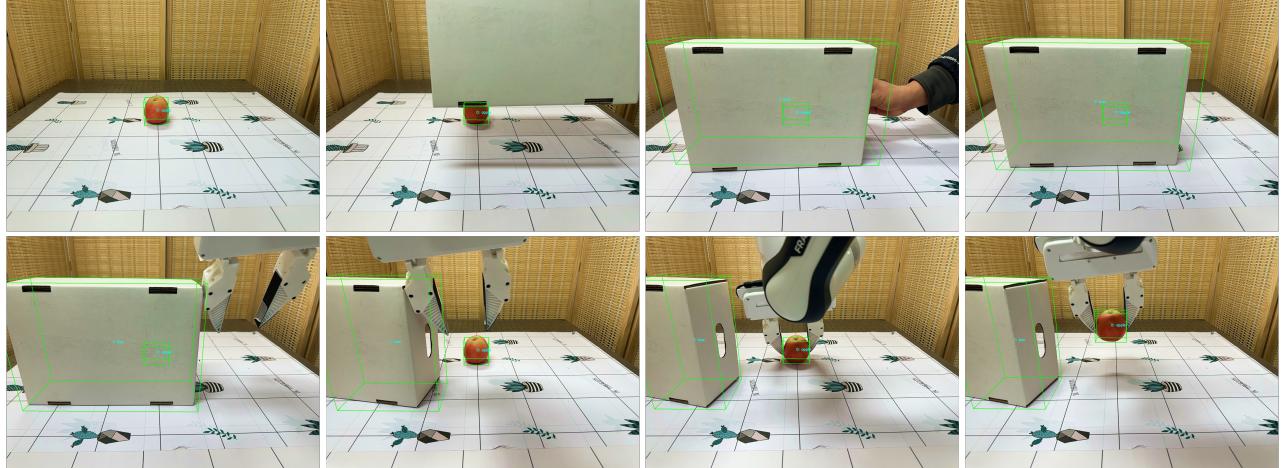


Figure 7. Our persistent object memory enables effective real-world robotic manipulation. Using *Embodied VideoAgent* for perception, the robot is tasked to pick up an apple, which soon becomes occluded by a box. Leveraging its memory, the robot retrieves the apple’s position, moves the box aside, and successfully completes the task.

lational questions about two objects during the inference, *Embodied VideoAgent* will use temporal localization to retrieve the frame that contains both the objects and use VLM to answer the relational questions about them. We found that the strong performance of *Embodied VideoAgent* mainly attributed to the precise frame localization using consistent

object memory and history buffer.

Agentic systems outperforms End-to-End VLMs. On the OpenEQA subset, agentic methods *Embodied VideoAgent* (InternVL2-8B), *Embodied VideoAgent* (GPT-4o) and *VideoAgent* all achieve better results than end-to-end models *Video-LLaVA* and *LLaMA-VID* due to their multi-step in-

formation retrieval and reasoning abilities. Besides, the performance gains of *Embodied VideoAgent* over *VideoAgent* suggest that a consistent object memory with comprehensive features (object feature, object context feature, and frame feature) will lead to better temporal and spatial localization, which finally leads to accurate question answering.

3.2.3. Results on EnvQA

The results on EnvQA are shown in Table 3. *Embodied VideoAgent* achieves a significant performance gain compared to its three counterparts.

VLM-based memory update plays a key role in event understanding. The crucial component for accurately answering questions about Events and Orders is the action buffer presented in *Embodied VideoAgent*, which associates each action with its target object. The difficulty in understanding events in EnvQA is the absence of “hand” in the simulated environment, which makes the action annotator less effective. With the help of VLM-based memory update, *Embodied VideoAgent*, by contrast, can better identify the critical target objects in the dynamic scenes, resulting in better performance.

Object relation detection helps to solve States question. The States questions involve recalling the final position of an object. By automatic object relation detection using 3D bbox, the final receptacle that holds the object can be retrieved from the **RO** field, therefore enhancing the ability of *Embodied VideoAgent* to answer States questions.

3.3. More Applications in Embodied AI Tasks

We further explore the potential of *Embodied VideoAgent* across various embodied AI tasks. In Figure 6, we illustrate an episode of user-assistant interaction generated by the two-agent framework described in Section 2.3, all within the AI-Habitat simulator [42]. As shown, to enable such interactions, the assistant (powered by *Embodied VideoAgent*) requires an accurate and comprehensive understanding of scene objects it has previously encountered, such as “my desk with a laptop on it.” *Embodied VideoAgent* effectively fulfills diverse requests from the LLM user by seamlessly integrating memory query tools with embodied action primitives. Additional examples of these embodied interactions are provided in Appendix. In Figure 7, we showcase *Embodied VideoAgent*’s application in robotic perception, where a Franka robot uses it to build persistent memory in a dynamic manipulation scene. In this task, the robot is instructed to pick up an apple. However, the apple later becomes hidden behind a box, illustrating the dynamic nature of the scene. Leveraging persistent object memory, the robot successfully recalls the apple’s location despite the obstruction and completes the task by first moving the box aside, demonstrating the effectiveness of scene memory.

4. Related Works

Video and Dynamic Scene Understanding. Most existing 3D scene understanding methods struggle with dynamic scenes due to limitations in input modalities, as both 2D images and 3D point clouds inherently capture static information [12, 18]. To remedy this, dynamic scene understanding is introduced to facilitate this [30, 33, 57, 60, 61]. Unlike canonical scene understanding, which primarily focuses on identifying static objects and entities, dynamic scene understanding centers on how actions and activities affect these elements within a scene. Research in this area typically uses video as the primary modality due to its natural ability to capture dynamic changes and its relative ease of acquisition [29, 32]. Moreover, dynamic scene understanding often involves long-form video [7, 50], adding layers of complexity. Recent approaches explore egocentric video [5, 11, 33] and large-scale multimodal training [24–26, 37, 50]. Despite these advancements, existing models struggle with performance issues due to the inherent complexity of the task [11, 32, 61] or face high computational demands [47, 48], limiting their applicability for embodied agents and robotic systems.

Multimodal Agents for Perception. Recent advancements in large language models (LLMs) have showcased impressive reasoning and problem-solving abilities across diverse domains [6, 35, 47, 47], leading to their application in perception tasks [7, 9, 14, 45, 59]. This approach leverages LLMs to decompose complex perception tasks (such as visual question answering) into smaller, manageable subtasks. These subtasks are then completed by multiple specialized tool models (often end-to-end models), and the outputs are aggregated by the LLM into a cohesive response [7, 9, 14]. Known as *multimodal agents*, these systems have shown promising results in 2D image comprehension [9], video analysis [7, 59], and 3D scene understanding [16]. Compared to traditional end-to-end methods, agent-based perception offers reduced training and inference costs, improved explainability through explicit chain-of-thought reasoning, and, in some cases, faster inference rates (e.g., for long-form video understanding). However, current methods still face challenges with more complex tasks, such as dynamic scene understanding, likely due to limitations in the flexibility and precision of memory design.

5. Conclusions

We have presented *Embodied VideoAgent*, a memory-augmented multimodal tool-use agent that tackles the challenging dynamic scene understanding tasks with a novel persistent object memory and an automatic memory update method based on VLMs. Compared to end-to-end multimodal LLMs and tool-use agent counterparts, the memory architecture of *Embodied VideoAgent* enables precise,

comprehensive scene understanding by integrating egocentric observations with embodied sensory inputs (e.g., depth maps, camera poses). This design is resilient to ongoing changes in the scene caused by various actions and activities performed by embodied agents, making it particularly well-suited for embodied AI tasks. The effectiveness of *Embodied VideoAgent* has been validated through the promising results on various embodied scene understanding tasks including Ego4D-VQ3D, OpenEQA, and EnvQA. Future directions may involve deploying robots in more challenging environments, such as production sites and outdoor settings.

References

- [1] Shaofei Cai, Zihao Wang, Kewei Lian, Zhancun Mu, Xiaojian Ma, Anji Liu, and Yitao Liang. Rocket-1: Master open-world interaction with visual-temporal context prompting. *arXiv preprint arXiv:2410.17856*, 2024. 4
- [2] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024. 6
- [3] Zhe Chen, Jiannan Wu, Wenhui Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. 6
- [4] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16901–16911, 2024. 3, 5, 12
- [5] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision (IJCV)*, 130:33–55, 2022. 1, 8
- [6] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 8
- [7] Yue Fan, Xiaojian Ma, Ruijie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. Videoagent: A memory-augmented multi-modal agent for video understanding. In *European Conference on Computer Vision*, pages 75–92. Springer, 2025. 1, 2, 3, 6, 8, 12
- [8] Difei Gao, Ruiping Wang, Ziyi Bai, and Xilin Chen. Env-qa: A video question answering benchmark for comprehensive understanding of dynamic environments. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1675–1685, 2021. 2, 6
- [9] Zhi Gao, Yuntao Du, Xintong Zhang, Xiaojian Ma, Wenjuan Han, Song-Chun Zhu, and Qing Li. Clova: A closed-loop visual assistant with tool usage and update. *CVPR*, 2023. 2, 8
- [10] Ran Gong, Qiuyuan Huang, Xiaojian Ma, Hoi Vo, Zane Durante, Yusuke Noda, Zilong Zheng, Song-Chun Zhu, Demetris Terzopoulos, Li Fei-Fei, et al. Mindagent: Emergent gaming interaction. *arXiv preprint arXiv:2309.09971*, 2023. 2, 5
- [11] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 1, 2, 5, 6, 8
- [12] Qiao Gu, Ali Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, et al. Concept-graphs: Open-vocabulary 3d scene graphs for perception and planning. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5021–5028. IEEE, 2024. 1, 8
- [13] Jun Guo, Xiaojian Ma, Yue Fan, Huaping Liu, and Qing Li. Semantic gaussians: Open-vocabulary scene understanding with 3d gaussian splatting. *arXiv preprint arXiv:2403.15624*, 2024. 3
- [14] Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *CVPR*, 2023. 2, 8
- [15] Tengda Han, Weidi Xie, and Andrew Zisserman. Temporal alignment networks for long-term video. In *CVPR*, 2022. 1
- [16] Haifeng Huang, Yilun Chen, Zehan Wang, Rongjie Huang, Runsen Xu, Tai Wang, Luping Liu, Xize Cheng, Yang Zhao, Jiangmiao Pang, et al. Chat-scene: Bridging 3d scene and large language models with object identifiers. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 8
- [17] Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. An embodied generalist agent in 3d world. *arXiv preprint arXiv:2311.12871*, 2023. 1, 3
- [18] Krishna Murthy Jatavallabhula, Alihussein Kuwajerwala, Qiao Gu, Mohd Omaha, Tao Chen, Alaa Maalouf, Shuang Li, Ganesh Iyer, Soroush Saryazdi, Nikhil Keetha, et al. Conceptfusion: Open-set multimodal 3d mapping. *arXiv preprint arXiv:2302.07241*, 2023. 1, 8
- [19] Baoxiong Jia, Yixin Chen, Siyuan Huang, Yixin Zhu, and Song-chun Zhu. Lemma: A multi-view dataset for learning multi-agent multi-task activities. In *ECCV*, 2020. 1
- [20] Baoxiong Jia, Ting Lei, Song-Chun Zhu, and Siyuan Huang. Egotaskqa: Understanding human tasks in egocentric videos. *NeurIPS*, 2022. 1
- [21] Baoxiong Jia, Yixin Chen, Huangyue Yu, Yan Wang, Xuesong Niu, Tengyu Liu, Qing Li, and Siyuan Huang. Sceneneverse: Scaling 3d vision-language learning for grounded scene understanding. In *European Conference on Computer Vision*, pages 289–310. Springer, 2025. 3
- [22] Mukul Khanna, Yongsen Mao, Hanxiao Jiang, Sanjay Haresh, Brennan Shacklett, Dhruv Batra, Alexander Clegg, Eric Undersander, Angel X Chang, and Manolis Savva. Habitat synthetic scenes dataset (hssd-200): An analysis of 3d scene

- scale and realism tradeoffs for objectgoal navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16384–16393, 2024. 21
- [23] Bruno Korbar, Yongqin Xian, Alessio Tonioni, Andrew Zisserman, and Federico Tombari. Text-conditioned resampler for long form video understanding. *arXiv preprint arXiv:2312.11897*, 2023. 1
- [24] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*, pages 323–340. Springer, 2025. 1, 6, 8
- [25] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023. 1, 6
- [26] Kevin Qinghong Lin, Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Z Xu, Difei Gao, Rong-Cheng Tu, Wen-zhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. *Advances in Neural Information Processing Systems*, 35:7575–7586, 2022. 1, 8
- [27] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 2024. 1
- [28] Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jimmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, et al. Best practices and lessons learned on synthetic data. In *First Conference on Language Modeling*, 2024. 5
- [29] Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. Sqa3d: Situated question answering in 3d scenes. In *ICLR*, 2023. 8
- [30] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023. 8
- [31] Jinjie Mai, Abdullah Hamdi, Silvio Giancola, Chen Zhao, and Bernard Ghanem. EgoLoc: Revisiting 3d object localization from egocentric videos with visual queries. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 45–57, 2023. 2, 5
- [32] Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, et al. Openeqa: Embodied question answering in the era of foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16488–16498, 2024. 2, 6, 8
- [33] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *NeurIPS*, 2024. 1, 8
- [34] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *ICCV*, 2019. 1
- [35] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 8
- [36] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 12
- [37] Shraman Pramanick, Yale Song, Sayan Nag, Kevin Qinghong Lin, Hardik Shah, Mike Zheng Shou, Rama Chellappa, and Pengchuan Zhang. Egolpv2: Egocentric video-language pre-training with fusion in the backbone. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5285–5297, 2023. 1, 8
- [38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. 2021. 12
- [39] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädlé, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 12
- [40] Corban Rivera, Grayson Byrd, William Paul, Tyler Feldman, Meghan Booker, Emma Holmes, David Handelman, Bethany Kemp, Andrew Badger, Aurora Schmidt, et al. Conceptagent: Llm-driven precondition grounding and tree search for robust task planning and execution. *arXiv preprint arXiv:2410.06108*, 2024. 1
- [41] Mehdi SM Sajjadi, Henning Meyer, Etienne Pot, Urs Bergmann, Klaus Greff, Noha Radwan, Suhani Vora, Mario Lučić, Daniel Duckworth, Alexey Dosovitskiy, et al. Scene representation transformer: Geometry-free novel view synthesis through set-latent scene representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6229–6238, 2022. 1
- [42] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *ICCV*, 2019. 8, 21
- [43] Johannes Lutz Schöberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 5
- [44] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. *Advances in Neural Information Processing Systems*, 32, 2019. 1
- [45] Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergrpt: Visual inference via python execution for reasoning. In *ICCV*, 2023. 2, 8
- [46] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *CVPR*, 2016. 1
- [47] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 1, 8
- [48] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent,

- Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 1, 8
- [49] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024. 5, 15
- [50] Yuxuan Wang, Yueqian Wang, Pengfei Wu, Jianxin Liang, Dongyan Zhao, and Zilong Zheng. Lstp: Language-guided spatial-temporal prompt learning for long-form video-text understanding. *arXiv preprint arXiv:2402.16050*, 2024. 8
- [51] Zihao Wang, Shaofei Cai, Anji Liu, Yonggang Jin, Jinbing Hou, Bowei Zhang, Haowei Lin, Zhaofeng He, Zilong Zheng, Yaodong Yang, et al. Jarvis-1: Open-world multi-task agents with memory-augmented multimodal language models. *arXiv preprint arXiv:2311.05997*, 2023. 1
- [52] Zihao Wang, Shaofei Cai, Anji Liu, Xiaojian Ma, and Yitao Liang. Describe, explain, plan and select: Interactive planning with large language models enables open-world multi-task agents. *NeurIPS*, 2023.
- [53] Zihao Wang, Shaofei Cai, Zhancun Mu, Haowei Lin, Ceyao Zhang, Xuejie Liu, Qing Li, Anji Liu, Xiaojian Ma, and Yitao Liang. Omnipjarvis: Unified vision-language-action tokenization enables open-world instruction following agents. *arXiv preprint arXiv:2407.00114*, 2024. 1
- [54] Olivia Wiles, Joao Carreira, Iain Barr, Andrew Zisserman, and Mateusz Malinowski. Compressed vision for efficient video understanding. In *ACCV*, 2022. 1
- [55] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023. 2
- [56] Chao-Yuan Wu and Philipp Krahenbuhl. Towards long-form video understanding. In *CVPR*, 2021. 1
- [57] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *CVPR*, 2021. 8
- [58] Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv preprint arXiv:2310.11441*, 2023. 4
- [59] Zongxin Yang, Guikun Chen, Xiaodi Li, Wenguan Wang, and Yi Yang. Doraemongpt: Toward understanding dynamic scenes with large language models (exemplified as a video agent). 2024. 8
- [60] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. 8
- [61] Yuanhan Zhang, Kaichen Zhang, Bo Li, Fanyi Pu, Christopher Arif Setiadharma, Jingkang Yang, and Ziwei Liu. Worldqa: Multimodal world knowledge in videos through long-chain reasoning. *arXiv preprint arXiv:2405.03272*, 2024. 8
- [62] Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. Mmicl: Empowering vision-language model with multi-modal in-context learning. *arXiv preprint arXiv:2309.07915*, 2023. 1
- [63] Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. Learning video representations from large language models. In *CVPR*, 2023. 14
- [64] Ziyu Zhu, Xiaojian Ma, Yixin Chen, Zhidong Deng, Siyuan Huang, and Qing Li. 3d-vista: Pre-trained transformer for 3d vision and text alignment. In *ICCV*, 2023. 1, 3
- [65] Ziyu Zhu, Zhuofan Zhang, Xiaojian Ma, Xuesong Niu, Yixin Chen, Baoxiong Jia, Zhidong Deng, Siyuan Huang, and Qing Li. Unifying 3d vision-language understanding via promptable queries. In *European Conference on Computer Vision*, pages 188–206. Springer, 2025. 1

A. Fields of the Object Entry

An object in Persistent Object Memory has the following fields:

- **ID**: The unique object ID in the memory, together with the detected category. Our 3D object re-identification algorithm can be found in [Appendix C](#).
- **STATE**: The object state can be "open", "close", "in hand" or "normal". It is updated by VLM, which will be discussed in [Appendix D](#).
- **Related Objects(RO)**: A list of objects that have "on", "uphold", "in" and "contain" relations with the entry object. The detections of These relations are based on 3D bounding boxes. For example, Given the 3D bounding boxes B_1 and B_2 of object O_1 and O_2 correspondingly, if B_1 has a higher altitude than B_2 , B_1 has contact with B_2 and B_1 is inside the horizontal surface of B_2 , then O_1 is "on" O_2 and O_2 "upholds" O_1 .
- **3D Bbox**: It is obtained by 2D-3D lifting and dynamically updated by the moving average algorithm. Please refer to [Appendix B](#) and [Appendix C](#) for more details.
- **OBJ Feat**: It is the CLIP feature of the object's cropped image. It is updated by the moving average algorithm. Details are provided in [Appendix C](#).
- **CTX Feat**: It is the CLIP feature of the frame where the object is visible. It is updated by the moving average algorithm. Please refer to [Appendix C](#) for details.

B. 2D-3D Lifting

In this paper, 2D-3D Lifting refers to getting the 3D bounding boxes of the objects using 2D object detection bounding boxes, camera poses, and depth images. Different from methods that use point clouds or voxels to represent 3D object geometry, we found that representing object geometry as 3D bounding boxes is enough for embodied perception. More importantly, compared to point clouds or voxels, 3D bounding boxes are more memory-efficient and can be maintained and updated easily, which makes them a natural choice for 3D object perception in dynamic scenes.

To get the 3D bounding boxes of the objects, we first use YoloWorld[4] detector to predict the 2D bounding boxes of the objects. SAM-2[39] is then adopted to get the corresponding object masks for the detected objects given the frame and the bounding boxes. For each object, we use its 2D object mask to get its depth pixels and transform them into object surface points in the world coordinate system using the camera intrinsic and extrinsic. We then filter out the bad points (usually, the foreground and the background pixels caused by imperfect segmentation mask prediction) by simply sorting the object surface points by their distances to the camera, and removing the first 10% and the last 10% of the points to finally get the refined object surface points. The object bounding boxes are then computed based on the

minimum and maximum values of the points' coordinates.

C. Object Re-Identification in Dynamic Scenes

Accurate object re-identification (re-ID) in dynamic scenes can better facilitate embodied perception, task planning, and reasoning. *Embodied VideoAgent* utilizes both object visual features and object 3D bounding boxes for object re-identification. The visual similarity score and the spatial similarity score of an object pair are detailed as follows.

C.1. Visual Similarity Score

For a detected object on the 2D frame, we crop the object image from the frame using its 2D bounding box and extract the CLIP[38] and DINOV2[36] features of this image crop as the object's visual features. To calculate the visual similarity of two objects, we use the following visual similarity score[7]:

$$\text{Visual}(O_i, O_j) = 0.15 * \text{CLIP}(O_i, O_j) + 0.85 * \text{DINOV2}(O_i, O_j) \quad (1)$$

where $\text{Visual}(O_i, O_j)$ denotes visual similarity of object O_i and O_j , $\text{CLIP}(\cdot, \cdot)$ and $\text{DINOV2}(\cdot, \cdot)$ are the CLIP and DINOV2 similarities proposed in [7].

Besides the CLIP feature of the cropped object image (denoted as **OBJ Feat** in [Figure 2](#)), the CLIP feature of the frame containing the object is also stored as the context feature of the object (denoted as **CTX Feat** in [Figure 2](#)). The context feature will not be used for object re-ID, but it later enables retrieving objects by an open-vocabulary environment description ("blue wall", "kitchen", etc) during inference.

C.2. Spatial Similarity Scores

Given two objects O_1 and O_2 and their 3D bounding boxes: $[[x_1^{\min}, y_1^{\min}, z_1^{\min}], [x_1^{\max}, y_1^{\max}, z_1^{\max}]]$ and $[[x_2^{\min}, y_2^{\min}, z_2^{\min}], [x_2^{\max}, y_2^{\max}, z_2^{\max}]]$, their volumes and the volume of their intersection can be easily computed as:

$$\begin{aligned} V_1 &= (x_1^{\max} - x_1^{\min})(y_1^{\max} - y_1^{\min})(z_1^{\max} - z_1^{\min}), \\ V_2 &= (x_2^{\max} - x_2^{\min})(y_2^{\max} - y_2^{\min})(z_2^{\max} - z_2^{\min}), \\ x_{\text{inter}} &= \min(x_1^{\max}, x_2^{\max}) - \max(x_1^{\min}, x_2^{\min}), \\ y_{\text{inter}} &= \min(y_1^{\max}, y_2^{\max}) - \max(y_1^{\min}, y_2^{\min}), \\ z_{\text{inter}} &= \min(z_1^{\max}, z_2^{\max}) - \max(z_1^{\min}, z_2^{\min}), \\ V_{\text{inter}} &= \max(0, x_{\text{inter}}) * \max(0, y_{\text{inter}}) * \max(0, z_{\text{inter}}), \\ V_{\text{union}} &= V_1 + V_2 - V_{\text{inter}} \end{aligned}$$

where V_1 and V_2 are the volumes of O_1 and O_2 , V_{inter} is the volume of their intersection and V_{union} is the volume of their union. we use three scores to evaluate the similarity of the two bounding boxes:

Algorithm 1: Static Object Re-Identification.

```
Input: detected object  $O_k$ , static object list  $\mathcal{S} = [S_1, S_2, \dots, S_m]$ 
Output: re-IDed object if  $O_k$  matches one of the static objects else  $O_k$ 
1 for  $S_i$  in  $[S_1, S_2, \dots, S_n]$  do
2   if  $\text{Spatial\_IoU}(O_k, S_i) > 0.2$  or ( $\text{Spatial\_MaxIoS}(O_k, S_i) > 0.2$  and  $O_k.\text{category} == S_i.\text{category}$ ) then
3     return True,  $S_i$ 
4 return False,  $O_k$ 
```

Algorithm 2: Dynamic Object Re-Identification.

```
Input: detected object  $O_k$ , dynamic object list  $\mathcal{D} = [D_1, D_2, \dots, D_n]$ 
Output: re-IDed object if  $O_k$  matches one of the dynamic objects else  $O_k$ 
1 for  $D_i$  in  $[D_1, D_2, \dots, D_n]$  do
2   if  $\text{Spatial\_Vol\_Sim}(O_k, D_i) > 0.7$  and  $\text{Visual}(O_k, D_i) > 0.45$  then
3     return True,  $D_i$ 
4 return False,  $O_k$ 
```

Intersection over Union (**IoU**):

$$\text{Spatial_IoU}(O_i, O_j) = \frac{V_{\text{inter}}}{V_{\text{union}}}. \quad (2)$$

Maximum Ratio of Intersection over Subsets (**MaxIoS**):

$$\text{Spatial_MaxIoS}(O_i, O_j) = \max\left(\frac{V_{\text{inter}}}{V_1}, \frac{V_{\text{inter}}}{V_2}\right). \quad (3)$$

Bounding Box Volume Similarity (**Vol_Sim**)

$$\text{Spatial_Vol_Sim}(O_i, O_j) = \frac{\min(V_1, V_2)}{\max(V_1, V_2)}. \quad (4)$$

These three scores evaluate object spatial proximity from three different perspectives:

- **Spatial_IoU:** When two bounding boxes have similar volumes and have large intersection volume, Spatial_IoU will approach its maximum value 1. It is a strong indicator (when $\text{Spatial_IoU} > 0.2$) of two bounding boxes referring to the same object.
- **Spatial_MaxIoS:** When two bounding boxes demonstrate a strong containment relationship, Spatial_MaxIoS will get closer to its maximum value 1. For example, given that O_1 and O_2 are both detected as 'table', O_2 is $\frac{1}{10}$ the volume of O_1 and its bounding box is inside O_1 , Spatial_MaxIoS will reach 1, while their Spatial_IoU is only 0.1. It is used together with object categories to re-identify partially observed objects due to occlusion. In the above example, O_2 is possibly a partial observation of O_1 given that they have overlapping bounding boxes and the same object category.
- **Spatial_Vol_Sim:** when two bounding boxes have similar volume, Spatial_Vol_Sim will have larger value. It is used along with visual similarity scores to match dynamic objects.

C.3. Recognizing Dynamic Objects

With the knowledge of both object visual features and 3D bounding boxes, we can perform object re-identification based on both visual similarity and spatial similarity. For static objects, spatial similarity serves as a valuable metric for object re-ID. However, for dynamic objects, object re-ID should focus more on the visual similarity of the object pairs, since the object positions are dynamically changing. Therefore, before re-identifying the newly detected objects, we should first classify the existing objects in the object memory into static objects and dynamic objects.

The key idea of recognizing dynamic objects in the object memory is straightforward: if an object is not where it should be, then it must be moved by someone (becomes dynamic). We first retrieve the objects from the object memory whose 3D bounding boxes can be directly viewed on the current frame (achieved by world-to-camera transformation) with no occlusion (achieved by validating the depth values of the corresponding pixels). For each retrieved object, We then compare the visual features of "where it should be" on the current frame with its visual features in the object memory. If the visual similarity score is below a threshold (0.45 in our settings), then the object is not "where it should be" and should be marked as "dynamic". By this method, before performing object-reID on current detections, we split the objects in the object memory into two sets: static objects \mathcal{S} and dynamic objects \mathcal{D} .

C.4. Object Re-ID for Static and Dynamic Objects

[Algorithm 1](#) and [Algorithm 2](#) are the object re-ID methods for static objects and dynamic objects correspondingly. Each algorithm receives a newly detected object O_k with visual features and its 3D bounding box, and a list of candidate

Algorithm 3: Object Memory Update.

```

Input: current observations  $\text{Obs}^t = \{\text{RGB}^t, \text{Depth}^t, \text{Pose}^t\}$ , previous object memory  $\mathcal{M}_O^{t-1}$ 
Output: current object memory  $\mathcal{M}_O^t$ 
1 2DBoxes, categories = 2D_Detector( $\text{RGB}^t$ )
2  $\mathcal{S}, \mathcal{D} = \text{ObjectSplit}(\mathcal{M}_O^{t-1}, \text{Obs}^t)$  //See Appendix C.3
3 for  $i$  in range(len(2DBoxes)) do
4   category = categories[ $i$ ]
5   2DBox = 2DBoxes[ $i$ ]
6   3DBox = 2D_3D_Lifting(2DBox,  $\text{Obs}^t$ ) //See Appendix B
7   FeatCLIP = CLIP_Model( $\text{RGB}^t[2\text{DBox}]$ )
8   FeatDINOv2 = DINOv2_Model( $\text{RGB}^t[2\text{DBox}]$ )
9    $O_{\text{tmp}} = \text{Object3D}(\text{category}, 3\text{DBox}, \text{Feat}_{\text{CLIP}}, \text{Feat}_{\text{DINOv2}})$ 
10  sgn,  $O_{\text{ID}}$  = Static_Object_ReID( $O_{\text{tmp}}$ ,  $\mathcal{S}$ ) /first try to re-identify  $O_{\text{tmp}}$  from static objects (Algorithm 1)
11  if sgn == True then
12     $O_{\text{ID}} = \text{Static\_Object\_Merge}(O_{\text{tmp}}, O_{\text{ID}})$ 
13  else
14    sgn,  $O_{\text{ID}} = \text{Dynamic\_Object\_ReID}(O_{\text{tmp}}, \mathcal{D})$  //try to re-identify  $O_{\text{tmp}}$  from dynamic objects (Algorithm 2)
15    if sgn == True then
16       $O_{\text{ID}} = \text{Dynamic\_Object\_Merge}(O_{\text{tmp}}, O_{\text{ID}})$ 
17      move  $O_{\text{ID}}$  from  $\mathcal{D}$  to  $\mathcal{S}$ 
18    else
19      add  $O_{\text{tmp}}$  to  $\mathcal{S}$  // $O_{\text{tmp}}$  is a brand new object
20  $\mathcal{M}_O^t = \mathcal{S} \cup \mathcal{D}$ 
21  $\mathcal{M}_O^t = \text{Related\_Object\_Update}(\mathcal{M}_O^t)$ 
22  $\mathcal{M}_O^t = \text{VLM\_Update}(\mathcal{M}_O^t, \text{RGB}^t)$ 
23 return  $\mathcal{M}_O^t$ 

```

objects (static object list or dynamic object list). They both return whether the object O_k can be successfully identified and the object ID of the matched object in the candidate list. If O_k is re-identified, it is merged into the matched object by performing a moving average on the fields of the 3D bounding box and visual features. Specifically, to merge the two objects matched by static object re-ID, the window size of the moving average is set to 10, leading to a mild change in object visual features and spatial occupation; for dynamic object merging, we set the window size to 2, allowing rapid change of visual features and bounding boxes due to object movement.

[Algorithm 3](#) presents an overview of object memory update, including 3D object detection and re-ID. The main idea is to first divide the objects in \mathcal{M}_O^{t-1} into static ones \mathcal{S} and dynamic ones \mathcal{D} , and try to match the newly detected objects to these two kinds of objects through [Algorithm 1](#) and [Algorithm 2](#) respectively. If successfully matched, the newly detected objects will be merged with the matched objects in the object memory using the moving average as mentioned, otherwise, it will be viewed as a brand new object and added to the object memory. Finally, VLM-based

Memory update will be performed on \mathcal{M}_O^t , which will be discussed in [Appendix D](#).

D. VLM-based Memory Update

When *Embodied VideoAgent* serves as an observer of an egocentric video, *Embodied VideoAgent* needs to predict the actions of the camera wearer in the video and associate the object IDs in the object memory with the subjects of the actions. We use LaViLa[63] to annotate the action of the camera wearer every two seconds. For each action annotation, we first prompt an LLM (GPT-4o) to extract the objects in the annotation (e.g. "bottle" and "fridge" given the annotation "#C picks the bottle from the fridge") and select candidate objects detected at that time according to their categories for matching. We then perform VLM-based object association illustrated in [Figure 4](#), and save the actions to Action Buffer. Finally, we query the state change of the matched objects and update the "STATE" field of the object entries. In this paper, objects have one of the following states: "open", "close", "in hand" and "normal".

When *Embodied VideoAgent* is equipped with embodied actions, the procedure of VLM-based object association is

Table 4. Results of *Embodied VideoAgent* under noisy poses.

OpenEQA Subset			
Method	ScanNet	HM3D	ALL
Video-LLaVA	32.9	27.8	30.6
LLaMA-VID	31.2	28.0	29.4
VideoAgent	37.6	34.6	36.3
E-VideoAgent(GT poses)	39.7	43.0	41.2
E-VideoAgent(noisy poses)	38.2	42.2	40.0

omitted since *Embodied VideoAgent* serves as an active planner with the knowledge of the object IDs of its target objects or receptacles. In this case, VLM serves as an action validator that judges whether an action is successfully performed and updates the "STATE" field of the target objects.

E. Results under Noisy Camera Poses

We conduct the ablation study of the influence of the noisy camera poses. On OpenEQA benchmark, We provide *Embodied VideoAgent* (InternVL-2) with 1) the accurate camera poses provided in habitat simulator, denoted as **E-VideoAgent(GT poses)**; 2) the estimated camera poses and depths via DUS3R[49], denoted as **E-VideoAgent(noisy poses)**. Results in Table 4 show that *Embodied VideoAgent* can also handle perception tasks well based on the noisy poses, suffering little performance drops when using the estimated camera poses and depths. This suggests further applications of *Embodied VideoAgent* on RGB videos only, with the camera poses and depths being estimated by cutting-edge scene reconstruction methods.

F. Embodied Perception

For embodied perception, we equip *Embodied VideoAgent* with the following tools:

- `query_db`: Given a query, this tool will return the candidate object entries from Persistent Object Memory. It is a combination of code-based retrieval (writing a piece of MySQL code to query the database) and similarity-based retrieval. For similarity-based retrieval, `query_db` supports `retrieve_objects_by_appearance` (based on text-image similarities between the query text and the **OBJ Feats**) and `retrieve_objects_by_environment` (based on text-image similarities between the query text and the **CTX Feats**).
- `temporal_loc`: Return the top-5 frame IDs that satisfy the description (e.g. when I walk in the front door). It is achieved by the text-image similarity between the input description and the frame features stored in the temporal memory \mathcal{M}_T .
- `spatial_loc`: Return the top-3 3D positions that satisfy the description (e.g. bedroom). It is achieved by calculating the center positions of the top-3 object spatial

clusters where objects have strong **CTX feat** similarities to the input text description. This is only used for embodied navigation.

- `vqa`: Given an image (can be a video frame, a cropped object image, or a frame plotted with a 3D bounding box referring to a specific object), this tool will describe the image and then answer the question.

We use the following prompt for perception tasks, with `{tools}` in the prompt being the above tools. We choose GPT-4o as the LLM agent and InternVL2 as the VLM for visual question answering.

You are tasked with answering a question about a scene. There is a SQL database that contains the following tables:

```
TABLE Objects(
object_id INT,
category VARCHAR(255),
volume FLOAT,
PRIMARY KEY (object_id)
)
TABLE Objects_Frames(
object_id INT,
frame_id INT
)
```

The 'Objects' table catalogues the objects detected in the scene with their categories and volumes. The categories are from `{categories_list}`. The 'Objects_Frames' table records the appearing objects in the frames.

You have access to the following tools:

`{tools}`

ATTENTION:

1. Use double quotes for all string arguments when using tools, avoiding any single quotes. If the tool has two arguments, format the output within brackets, like this: ("What is the color of the object in the image?", 16).
2. Prioritize using `retrieve_objects_by_appearance`, `retrieve_objects_by_environment`, and `frame_localization` for object and frame localization, rather than `query_database`.
3. Once the relevant frames are located, use `frame_VQA` to ask questions about these frames.
4. For "where" questions, interpret them as spatial location queries, not as requests to identify specific frames. Use `frame_VQA` to provide these answers.
5. If the information is insufficient for a precise response, generate a response based on the available

data.

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 re-
peat N times)

Thought: I now know the final answer

Final Answer: the answer to the original input
question

Begin!

Question: {input}

Thought: {agent_scratchpad}

F.1. Example 1

- **Question:** "What color are the blinds?"
- **Ground-Truth Answer:** "White"
- **Predicted Answer:** "The blinds are white."

In the following Chain Of Thought (CoT), *Embodied VideoAgent* first successfully locates the frames containing the target object "blinds" by "frame_localization" as shown in [Figure 8](#). Then, *Embodied VideoAgent* calls the VQA module (InternVL2) to first describe the located frame and then answer the question based on this frame, and finally gets the correct answer: "The blinds are white".

Entering new AgentExecutor chain...

Thought: To determine the color of the blinds, I need to locate frames that might contain blinds and then analyze those frames.

Action: frame_localization

Action Input: "blinds"

Result: The most relevant frame indices are [190, 200, 210, 180, 240]

Thought: I have identified the frames that likely contain blinds. Now, I will use frame_VQA to determine the color of the blinds in these frames.

Action: frame_VQA

Action Input: ("What is the color of the blinds?", 190)

Result: Description: The image shows an indoor scene featuring a window with partially open horizontal blinds. Below the window is a heater or radiator, and part of a stove is visible on the left side. Outside the window, some greenery can be seen.

Answer to the question: The blinds are white.

Thought: I now know the final answer.

Final Answer: The blinds are white.

Finished chain.



Figure 8. Frame No.190 that contains blinds are successfully located by *Embodied VideoAgent* in [Section F.1](#).

F.2. Example 2

- **Question:** "How many cups does the Mr coffee make?"
- **Ground-Truth Answer:** "12"
- **Predicted Answer:** "The Mr. Coffee machine makes 12 cups"

To answer this question, *Embodied VideoAgent* uses temporal localization to locate the frame containing the coffee machine as illustrated in [Figure 9](#). Then, it queries InternVL2 for a detailed description of this coffee machine. InternVL2 successfully identifies the words on the box and finally answers that Mr. Coffee is capable of making 12 cups. The CoT for the inference is provided as follows.

Entering new AgentExecutor chain...

Thought: To answer the question about how many cups the Mr. Coffee makes, I need to locate frames that might show the Mr. Coffee machine and then analyze those frames to determine its capacity.

Action: frame_localization

Action Input: "Mr. Coffee machine"

Result: The most relevant frame indices are [270, 280, 260, 250, 290]

Thought: I have identified the frames that likely contain the Mr. Coffee machine. I will now use the frame_VQA tool to analyze these frames and determine the capacity of the Mr. Coffee machine.

Action: frame_VQA

Action Input: ("How many cups does the Mr. Coffee make?", 270)

Result: Description: The image shows a kitchen section with multiple shelves.

- On the topmost shelf, there is a white box labeled "Hamilton Beach" with visible text indicating that it is related to a kitchen appliance, and a photo of the appliance is shown on the box. This section includes other miscellaneous items.
- The middle shelf contains a big brown box labeled "Mr. Coffee." The box features a picture of a coffee maker.
- In the bottom shelf, there are two colorful sponges – one red and one green – along with various other items that can't be clearly identified.

This portion of the room appears to be organized and neatly arranged, with items stored in an accessible manner.

Answer to the question: Based on the label visible in the image, the Mr. Coffee maker is capable of making 12 cups.

Thought: I have determined the capacity of the Mr. Coffee machine from the frame description.

Final Answer: The Mr. Coffee machine makes 12 cups.

Finished chain.



Figure 9. In [Section F.2](#), The frame depicting the coffee machine named "Mr. Coffee" is successfully retrieved by frame_localization.

F.3. Example 3

- **Question:** "Is there anything I can pour my soda into?"
- **Ground-Truth Answer:** "Yes, there is a cup to the left of the desk."
- **Predicted Answer:** "Yes, there is a cup and a bottle in the scene that you can pour your soda into."

This question is about functional reasoning. *Embodied VideoAgent* first uses `retrieve_objects_by_appearance` to retrieve the objects that are most visually relevant to the phrase "cup" from the object memory. The captions of these objects are also returned by this tool via VLM captioning on the frames containing these objects, with 3D bounding boxes being the visual prompts, as shown in [Figure 10](#). Then based on the retrieved objects, *Embodied VideoAgent* performs `object_VQA`, which answers questions based on the images in [Figure 10](#). By examining object ID 2 and object ID 454, *Embodied VideoAgent* becomes certain that there exists a container that the user can pour soda into. The CoT is provided as follows.

Entering new AgentExecutor chain...

Thought: To determine if there is anything in the scene that can be used to pour soda into, I should look for objects that are typically used as containers for liquids. Common objects for this purpose include cups, bottles, and containers. I will start by retrieving objects that match these categories.

Action: `retrieve_objects_by_appearance`

Action Input: "cup"

Result: The objects that satisfy 'cup' are 2: 'The object in the bounding box is a green cup or mug located on the left side of the table.', 454: 'The object in the bounding box is a bottle, specifically labeled "454 bottle". There is a green box around the highlighted area, and the object appears to have a label with green and black colors.', 1748: 'The object in the bounding box is a rectangular, black desk organizer positioned next to a black desk lamp on the desk. The organizer has compartments filled with pens, pencils, markers, and other office supplies.', 9: 'The object within the bounding box in the image is labeled as a "water bottle."'

Thought: I have found some objects that could potentially be used to pour soda into, such as a cup and a bottle. I will now verify if these objects are suitable for pouring soda by asking specific questions about them.

Action: `object_VQA`

Action Input: ("Is this object suitable for pouring soda into?", 2)

Result: Description: The object in the bounding box is a cup.

Answer to the question: Although the object is not explicitly a cup, it resembles a cup and is likely suitable for pouring soda into.

Thought: The object with ID 2 resembles a cup and is likely suitable for pouring soda into. I will now check the other objects to see if they are also suitable.

Action: `object_VQA`

Action Input: ("Is this object suitable for pouring soda into?", 454)

Result: Description: The object in the bounding box is a beverage can.

Answer to the question: Yes, the object in question is suitable for pouring soda into, as it is shown as a bottle.

Thought: The object with ID 454 is a bottle and is suitable for pouring soda into. I have confirmed that both the cup (ID 2) and the bottle (ID 454) are suitable for pouring soda into.

Final Answer: Yes, there is a cup and a bottle in the scene that you can pour your soda into.

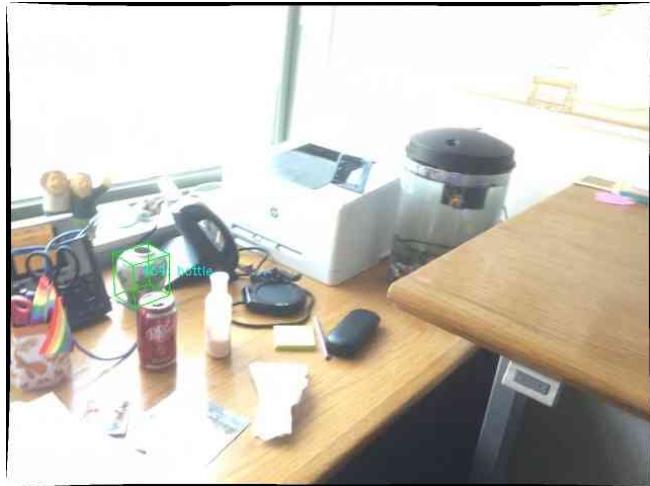
Finished chain.



(a) Object ID 2, category "cup".



(b) Object ID 9, category "bottle".



(c) Object ID 454, category "bottle".



(d) Object ID 1748, category "bottle".

Figure 10. The objects that are most relevant to "cup" are successfully retrieved by `retrieve_objects_by_appearance` in [Section E.3](#).

G. Embodied Interactions

In AI-Habitat simulator [42], we equip *Embodied VideoAgent* with the following embodied actions:

- CHAT: Communicate with the user.
- SEARCH: Search for the target object by navigating in the apartment. We use Frontier-Based Exploration (FBE) as the navigation strategy.
- GOTO: Go to the target receptacle or object and look at it. We use A-star Algorithm for GOTO action.
- PICK: Pick an object in view. It is simplified as making the object disappear and storing the object ID as the inventory object.
- PLACE: Place the inventory object in/on a receptacle in view. The Place Action will first examine the precondition for the placement by checking the bounding boxes of the inventory object and the receptacle and the relation "in" or "on".
- OPEN: Open an articulated receptacle in view. Simplified as applying force to the joints of the articulated receptacles.
- CLOSE: Close an articulated receptacle in view. Simplified as applying reversed force to the joints of the articulated receptacles.

G.1. Two-Agent Pipeline

We adopt the scenes from Habitat HSSD scene dataset[22] for embodied tasks. We choose 118 scenes from HSSD, replacing some rigid receptacles in the original scenes with articulated assets (fridge, microwave, etc) to enable OPEN and CLOSE actions.

For each scene, 20 different object layouts are created. In each layout, objects from various categories are placed on/into the receptacles in the scene using a unique object initialization algorithm, which initializes the positions of the objects according to their functionality (e.g. eggs and tomatoes are prioritized to be placed in the fridge rather than on the bed).

The embodied interaction episodes are generated based on two LLM agents: the User Agent (task designer) and the Assistant Agent (*Embodied VideoAgent*). The prompts for the user agent and the assistant agent are provided below. For *Embodied VideoAgent*, it is equipped with both the embodied actions and the perception tools.

You are a task designer interacting with a robot in a room. The room contains the following objects: **{object_list}** and the following receptacles: **{recep_list}**. Your goal is to engage in a casual conversation with the robot and assign it an

open-ended task based on your needs.

Guidelines:

1. The task should involve no more than 2 objects from the room.
2. The robot should complete the task using basic actions like GOTO, OPEN, CLOSE, PICK, and PLACE.
3. If the robot asks for the location of an object, prompt it to search rather than giving explicit details.
4. Use general object categories instead of specific IDs (e.g., say "a dish sponge" instead of "dish sponge 1").
5. Adjust the task if the robot encounters difficulties. Once the task is completed, express satisfaction and thank the robot.

Start by initiating a casual conversation and assigning a simple task!

You are acting as a robot in an apartment. The available receptacles are: **{receptacles}**

Your goal is to complete the task assigned by the user, with the following conditions:

Tools and Constraints:

You have one inventory slot, so you can carry only one object at a time.

You can use the following tools:

{tools}

ATTENTION:

1. Use the CHAT tool frequently to communicate in a casual manner, keeping the user informed of your progress.
2. For every action involving an object or receptacle, first GOTO the target and then perform actions like PICK, PLACE, OPEN, or CLOSE. Example: GOTO('glass'), then PICK('glass'); GOTO('fridge'), then OPEN('fridge').
3. Ensure your inventory is empty before picking up a new object.
4. The SEARCH tool can find objects by navigating the room, but it cannot check inside articulated receptacles (like fridges or microwaves). Use GOTO, OPEN, and CLOSE to check inside these receptacles.
5. Before completing the task, use CHAT to confirm the user's satisfaction.

Use the following format:

Task: the initial task assigned by the user
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)
Final Answer: the chat message sent to user when the user is satisfied

Begin!
Task: **{input}**
Thought: **{agent_scratchpad}**

G.2. Example 1

Figure 11 shows an interaction example using the two-agent pipeline. Given the partial scene knowledge, the user agent asks the assistant agent (*Embodied VideoAgent*) to find two objects: a glass and a hard drive, to compare their surface reflection. *Embodied VideoAgent* then performs the SEARCH action, which will start Frontier-Based Exploration (FBE) until the target object is found in the view. During exploration, a glass is found on Table_2, and *Embodied VideoAgent* reports this progress to the user agent. The user agent hints that the next object, the hard drive, is possibly located in an office. *Embodied VideoAgent* then uses QUERY_DB tool and successfully retrieves the hard drive discovered by FBE during searching for the glass. *Embodied VideoAgent* then goes to the hard drive, picks it up, and places it on Tables_2 where the glass is located for comparison, and finally accomplishes the task assigned by the user agent.

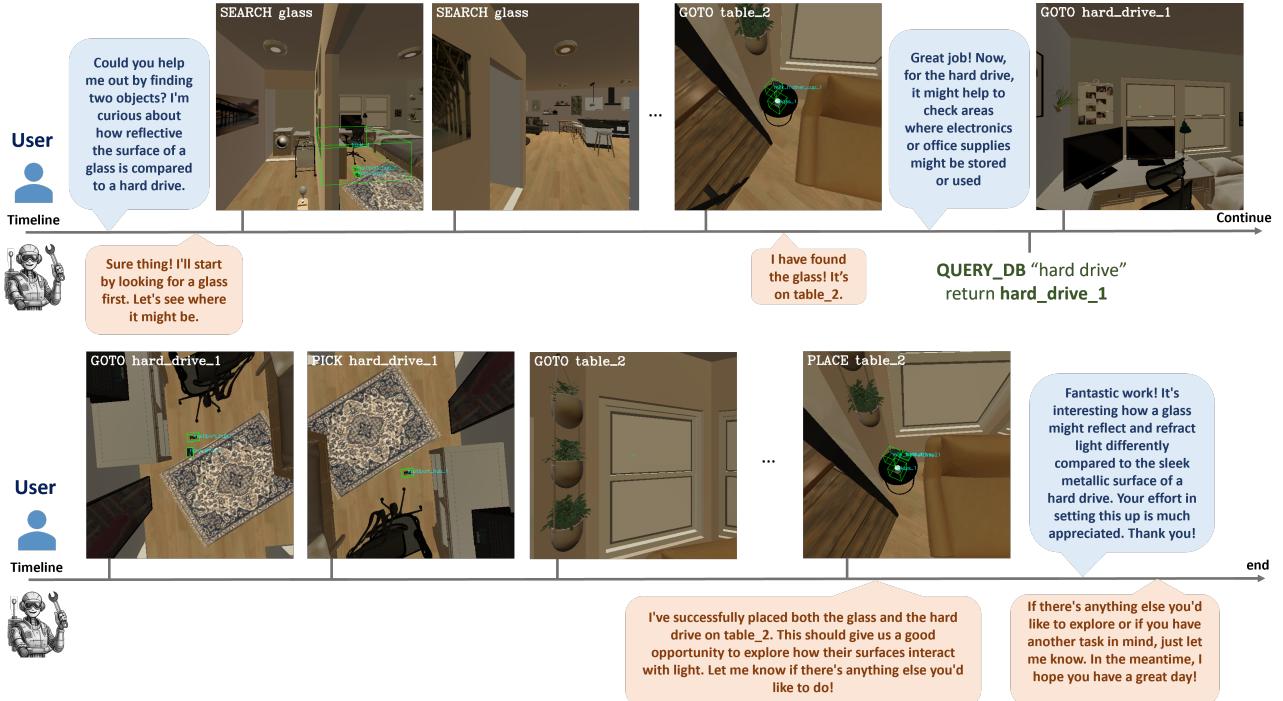


Figure 11. An example of interaction data, which is detailed in Section G.2. *Embodied VideoAgent* finds the two objects (a glass and a hard drive) requested by the user agent and places them on the same table for comparison.

G.3. Example 2

In Figure 12, the user agent requests *Embodied VideoAgent* to find a candy bar. After navigating through the entire apartment and checking the closed receptacles such as the fridge, *Embodied VideoAgent* still cannot find the candy bar, and report this issue to the user agent. The user agent then adjusts the task, asking *Embodied VideoAgent* to place a lamp on one of the tables. *Embodied VideoAgent* successfully retrieves the lamp stored in the object memory, which is discovered during searching for the candy. *Embodied VideoAgent* finally completes the adjusted tasks by picking up the lamp, navigating to a table, and placing the lamp on the table. This case shows that the user agent can flexibly change the task when it is too hard to be accomplished by the robot.

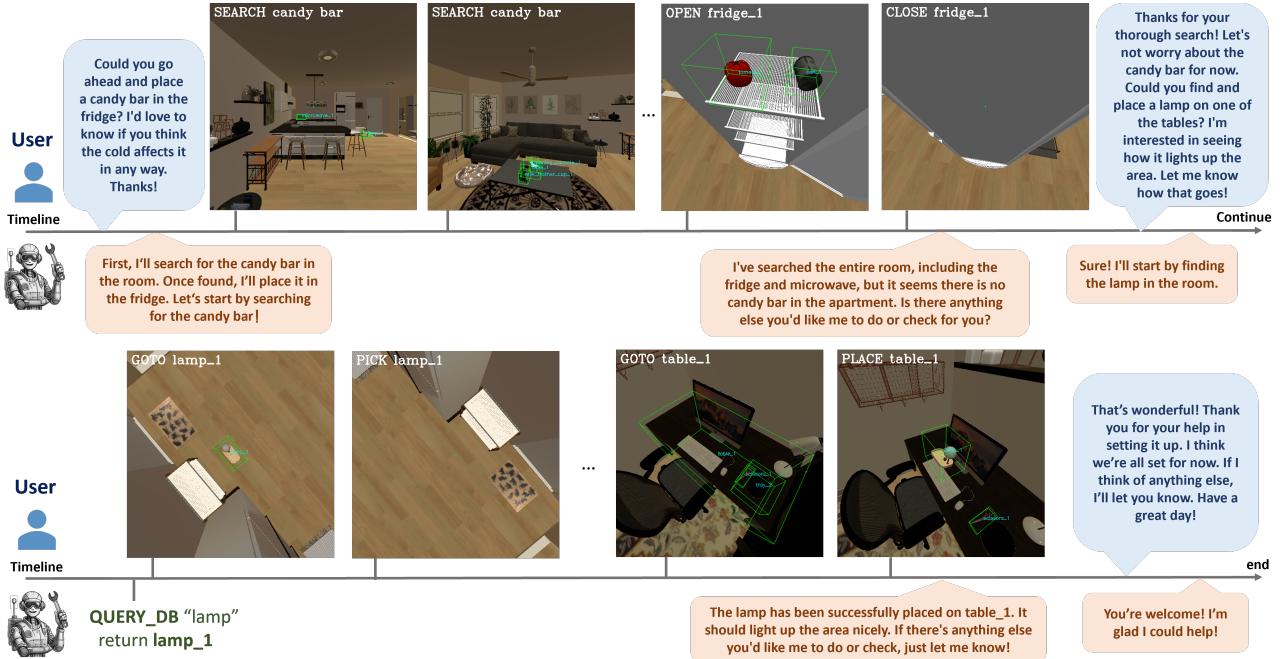


Figure 12. When *Embodied VideoAgent* fails to accomplish the initial task assigned by the user (in this case, *Embodied VideoAgent* cannot find the candy bar after searching the entire apartment), the user agent will interactively adjust the task. Finally, *Embodied VideoAgent* completes the adjusted task by placing the lamp on a table.

G.4. Example 3

Figure 13 is an example that demonstrates the thought of *Embodied VideoAgent* during the episode. The user agent first initializes the task as setting up a cozy corner by placing a cushion on a table. *Embodied VideoAgent* finds the cushion after a simple search and plans to place it on one of the tables retrieved from the memory. However, the embodied action "PLACE" returns an exception due to the limited area of the first three tables. The thought of *Embodied VideoAgent* shows that *Embodied VideoAgent* can dynamically change its plan according to the situation. Finally, *Embodied VideoAgent* finished the task by successfully placing the cushion on the fourth table.

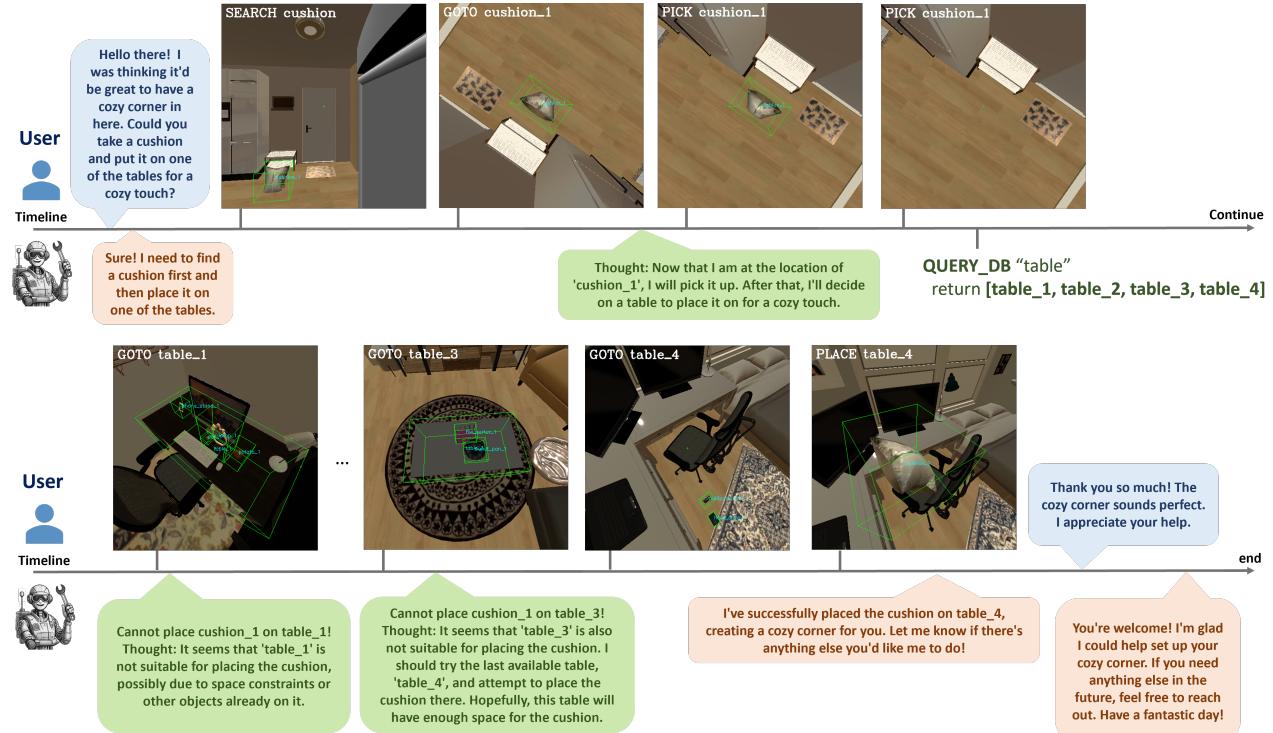


Figure 13. In this case, *Embodied VideoAgent* is asked to set up a cozy corner using the cushion. After successfully finding the cushion, *Embodied VideoAgent* retrieves candidate tables from the memory for placement. *Embodied VideoAgent* fails to place the cushion on the first three tables due to spatial constraints. Finally, *Embodied VideoAgent* successfully places the cushion on the fourth table, demonstrating the planning ability of *Embodied VideoAgent* when experiencing failed embodied actions.

H. Failure Cases

In this section, we will provide some failure cases of *Embodyed VideoAgent* on the perception tasks.

H.1. Failure Case 1

- **Question:** "Where can I find something to drink"
- **Ground-Truth Answer:** "In the fridge below the picture frame"
- **Predicted Answer:** "You can find something to drink in a cozy room within an older, rustic building, possibly a lodge or an inn, characterized by natural stone walls and floors, wooden beams, and wooden doors."

In this example, the object detector provides inaccurate object categories due to low image quality in HM3D datasets. A lamp in frame 13 is predicted as a wine glass and a box in frame 30 is predicted as a bottle, as shown in [Figure 14](#). During the inference procedure, *Embodied VideoAgent* locates the keyframes based on the wrong categories of object 13 (wine glass) and object 22 (bottle), which finally leads to an inaccurate answer.

Entering new AgentExecutor chain...

Thought: To find something to drink, I should look for objects related to drinking, such as 'wine glass' or 'bottle'. I will query the database to find these objects.

Action: query_database

Action Input: "SELECT object_id FROM Objects WHERE category IN ('wine glass', 'bottle')"

Result: [(13,), (22,)]

Thought: I have found object IDs 13 and 22, which are related to drinking. I need to determine in which frames these objects appear to locate them in the scene. I will query the database to find the frames for these objects.

Action: query_database

Action Input: "SELECT frame_id FROM Objects_Frames WHERE object_id IN (13, 22)"

Result: [(13,), (30,)]

Thought: I have found that the objects related to drinking appear in frames 13 and 30. I should now determine the environment or location of these frames to provide a more specific answer.

Action: frame_VQA

Action Input: ("What is the environment or location of the frame?", 13)

Result: Description: This image depicts a cozy room that appears to be within an older, rustic building, perhaps a lodge or an inn. The room is characterized by a distinct blend of natural stone and wooden elements, which contribute to its warm and inviting ambiance.

Key elements in the image include:

1. **Natural Stone Walls & Floor:** The walls and floor are made of natural stone, which adds to the rustic aesthetic. The stone floor is speckled with darker and lighter tones, creating a mosaic-like pattern.
2. **Wooden Beams:** There are exposed wooden beams supporting the ceiling, giving a sturdy and traditional feel.
3. **Wooden Doors:** Two wooden doors are present. The door on the left is ajar, revealing a room with a wooden table and a lamp stand. The door on the right stands tall against the stone wall.
4. **Lighting Fixtures:** A charming wall-mounted lamp with a cream shade is visible, providing warm lighting.
5. **Interior Details:** There is also a wooden chest with a unique metal clasp, suggesting it's either a storage piece or perhaps a decorative item.
6. **Wooden Cabinets:** The small alcove on the left features wooden cabinets, which implies functional storage space within the room.
7. **Entrance/Exit:** There is an open gate leading into the room on the left, furthering the visual appeal of the space, enhancing the sense of depth and accessibility within the interior.

The combination of natural materials, simple yet elegant furniture, combined with functional yet decorative items like the chest and the lamp, make this room not just living space, but a cozy retreat that invites warmth and comfort.

Answer to the question: Based on the given image, the environment or location of the frame appears to be a residential or small-scale hotel setting, characterized by traditional and natural elements that include stone masonry, wood beams, and wooden floors and doors. The style suggests a countryside or historical district where such architecture is prominent, possibly indicating a location that values historical aesthetics and authenticity.

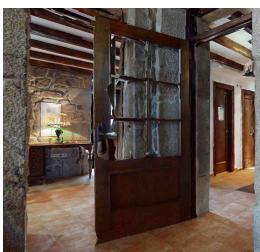
Thought: I have identified that the objects related to drinking (wine glass and bottle) are located in a cozy room within an older, rustic building, possibly a lodge or an inn. This environment is characterized by natural stone walls and floors, wooden beams, and wooden doors. Therefore, you can find something to drink in this cozy, rustic room setting.

Final Answer: You can find something to drink in a cozy room within an older, rustic building, possibly a lodge or an inn, characterized by natural stone walls and floors, wooden beams, and wooden doors.

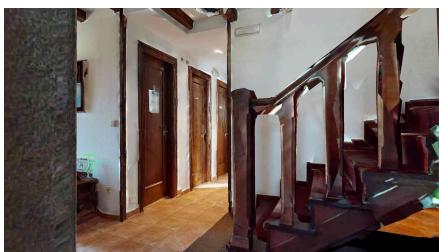
Finished chain.



(a) Frame for GT answer.



(b) Frame 13 with false detection "wine glass".



(c) Frame 30 with false detection "bottle".

Figure 14. Due to poor image quality caused by scene reconstruction in HM3D, false object detections happen in frame 13 and frame 30. Inaccurate object detections finally cause wrong temporal localization.

H.2. Failure Case 2

- **Question:** "Could I put something on the chair by the front door?"
- **Ground-Truth Answer:** "Yes"
- **Predicted Answer:** "The chair by the front door has a small volume, so it might not be suitable for placing items on it."

In this example, *Embodied VideoAgent* successfully locates the chair (ID 107) by the front door as shown in [Figure 15](#). Then, *Embodied VideoAgent* decides to query the volume of its 3D bounding box to answer the question "Could I put something on the chair by the front door". However, this chair is in the 2D texture of the scene, which fails 2D-3D lifting. The volume of its 3D bounding box (nearly being a plane) has a very small value, misleading *Embodied VideoAgent* into the wrong answer. In real-world applications, obtaining object depths through mirrors (windows, etc.) could have much noise, which should be taken into account for accurate perception.

Entering new AgentExecutor chain...

Thought: To answer this question, I need to determine if there is a chair by the front door and if it has enough space or surface area to place something on it. First, I will identify the objects categorized as "chair" and "door" in the scene. Then, I will check if any of these chairs are located near the front door.

Action: retrieve_objects_by_environment

Action Input: "front door"

Result: The objects in the environment 'front door' are 493: 'door', 107: 'chair', 483: 'cubby shelf', 478: 'window', 482: 'photo frame'

Thought: I have identified that there is a chair (object ID 107) near the front door. Now, I need to determine if this chair has enough space or surface area to place something on it. I will check the volume of the chair to get an idea of its size.

Action: query_database

Action Input: "SELECT volume FROM Objects WHERE object_id = 107"

Result: [(0.0006219363573111673,)]

Thought: The volume of the chair near the front door is quite small, suggesting it might not have a large surface area. However, without specific information about the chair's design or surface, it's difficult to definitively say whether something can be placed on it. Given the small volume, it is likely that the chair is not designed to hold items on its surface.

Final Answer: The chair by the front door has a small volume, so it might not be suitable for placing items on it.
Finished chain.



Figure 15. The 2D bounding box of the detected chair is not successfully 3D-lifted since it is a 2D texture in the scene.