This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

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 Chen[1], Qing Li[1†]
[1]State Key Laboratory of General Artificial Intelligence, BIGAI, Beijing, China
[2]University of Science and Technology of China [3]Tsinghua University [4]Peking 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* 6.5% *on Ego4D-VQ3D,* 2.6% *on OpenEQA, and* 15.3% *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 [16, 52–54]. In this paper, we investigate approaching this problem using egocentric observations [3, 5, 9, 19, 25, 33, 38], 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*.) [16, 42, 44, 69]; 2) handling dynamic environments as actions and activities might be performed by embodied agents themselves and other co-habited characters [3, 9, 33]; 3) maintaining a persistent memory about the scene that allows frequent update over time [10, 17, 41]. However, existing efforts on this front mostly adopt end-to-end pretrained multimodal large models (MLMs) [16, 23, 24, 27, 63, 67, 68]. Their capabilities of handling long-form videos and embodied sensory observations have been questioned by several prior studies [22, 34, 47, 57], especially when the scene being depicted is highly volatile with complex events and spatial-temporal dependencies [13, 18, 19]. Some sophisticated MLMs coming out recently have attained great progress in understanding long-form videos and the underlying embodied scenes [48, 49, 55], 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 approaches – *multimodal agents* [5, 7, 12, 45, 56]. These

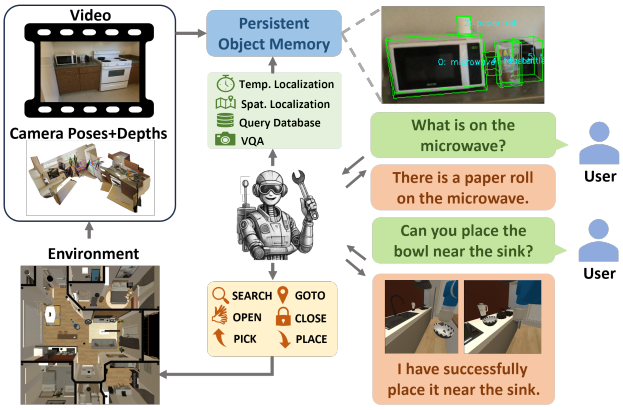


Figure 1. *Embodied VideoAgent* is a multimodal agent that 1) builds scene memory from both egocentric video and embodied sensory input; 2) utilizes multiple tools to query this memory; 3) activates embodied action primitives to interact with the environments, effectively fulfills various user requests.

*Equal contribution. †Corresponding authors.

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 [5, 7, 45]. However, extending these methods to dynamic 3D scene understanding is **non-trivial**. The system must develop a comprehensive yet precise understanding of objects within a constantly evolving environment, where various actions and activities performed by embodied agents and other entities induce continuous changes. Our initial explorations (see Section 5) with systems like VideoAgent [5] reveal that relying solely on visual frames for scene memory fails to construct a coherent spatial representation of the scene and its objects. Additionally, such an approach struggles to capture dynamic changes effectively, resulting in suboptimal embodied perception.

To this end, we propose *Embodied VideoAgent*, a multimodal agent that builds embodied perception from video frames and embodied sensors (e.g., depth and pose), as illustrated in Figure 1. Key components include: 1) a **persistent object memory** integrating egocentric video and sensor input to accurately track and re-identify 3D objects with attributes; 2) a **VLM-based update mechanism** that detects object changes during actions and updates memory accordingly. Together, these enable spatial and visual understanding of dynamic 3D scenes.

Beyond passive observation (*e.g*., on smart glasses), *Embodied VideoAgent* supports active control with action primitives like GOTO, PICK, and PLACE, forming a closed loop of perception and action. To evaluate *Embodied VideoAgent*, we design a **two-agent framework** in Habitat-Sim [43, 46], where an LLM user agent issues requests and *Embodied VideoAgent* acts to complete them, providing real-time feedback. A VLM further evaluates task success and data quality from visuals and dialogue, supporting scalable embodied data collection.

Our contributions can be summarized as follows:

- We introduce a persistent object memory to track and maintain the states of dynamic 3D objects, leveraging a unique 3D object re-identification (re-ID) algorithm. To ensure accurate updates, we design a VLM-based memory update mechanism that modifies object states based on their associated actions.
- Building on the resulting agent, *Embodied VideoAgent*, we develop a two-agent framework that facilitates embodied user-assistant interactions. In this framework, the user issues embodied tasks, while *Embodied VideoAgent* actively explores the scene to progressively complete them.
- On various scene understanding tasks, *Embodied VideoAgent* achieves improvements of 6.5% on VQ3D [9, 31], 2.6% on OpenEQA [32], and 15.3% on EnvQA [6] over canonical end-to-end multimodal LLMs and other multimodal agents. Additionally, the multi-agent framework highlights the effectiveness of *Embodied VideoAgent* in embodied perception and task planning, enabling the generation of high-quality embodied episodes.

2. Related Work

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 [10, 17]. To remedy this, dynamic scene understanding is introduced to facilitate this [14, 30, 33, 37, 58, 61, 62, 65]. 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 egocentric video [3, 9, 33] 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 videos[5, 51], adding layers of complexity. Recent approaches explore large-scale multimodal training [23–25, 38, 51]. However, existing models struggle with performance issues due to the inherent complexity of the task [9, 32, 62] or face high computational demands [48, 49], 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 [4, 35, 48, 48], leading to their application in perception tasks [5, 7, 12, 45, 60]. Multimodal agents leverage 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 [5, 7, 12]. Multimodal agents have shown promising results in 2D image comprehension [7], video analysis [5, 60], and 3D scene understanding [15]. Compared to traditional end-to-end methods, agent-based perception offers reduced training and inference costs and improved explainability through explicit chain-of-thought reasoning. However, they still face challenges with dynamic scene understanding due to the lack of a persistent object memory that can reflect dynamic scenes.

3. Embodied VideoAgent

Before introducing *Embodied VideoAgent*, we first examine whether agents based on 2D visual frames only, specifically VideoAgent [5], can be directly applied to embodied per-

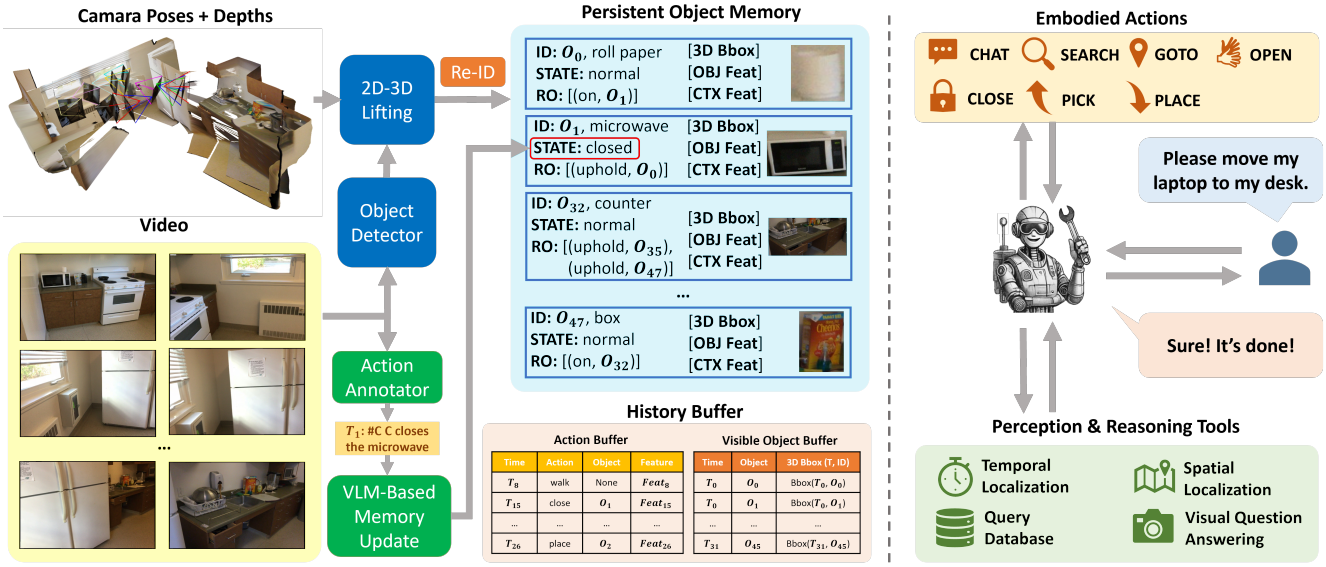


Figure 2. An overview of *Embodied VideoAgent*. Left: We first translate the egocentric video and embodied sensory input (depth maps and camera poses) into structured representations: *persistent object memory* and history buffer. While the memory can be updated using VLM to support dynamic scenes where 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.

ception in dynamic scenes. VideoAgent builds a temporal memory of video descriptions and an object memory that tracks occurrences of objects and people from 2D video frames. Imagine wearing smart glasses powered by VideoAgent: you place a cup on the table, turn to wash dishes, and return to find the cup missing. You may infer that a family member took it, but VideoAgent may fail to detect this change, as no new object is identified and the action of someone picking up the cup is not captured in the frames.

To address this, 3D geometry from embodied sensor inputs can be integrated with 2D visual frames to enhance perception. The aforementioned issue can be resolved by continuously comparing the memorized 3D bounding box of the cup with the current perceived 3D geometry. If the perceived 3D geometry and 2D visual features do not align with the memorized cup, then the cup is marked as 'dynamic' indicating that it has been moved. This approach improves the agent's spatial intelligence and its ability to detect changes in dynamic environments.

Embodied VideoAgent is therefore designed with such enhanced object memory for embodied perception, as shown in Figure 2. Additionally, a history buffer saves the past actions as well as the historical frames with appearing object IDs. For an embodied task, *Embodied VideoAgent* enters a perception-decision-action loop, iterating between perception tools (interacting with memory) and action primitives (interacting with the environment) until the task is completed. The details of *Embodied VideoAgent* are as follows.

3.1. Persistent Object Memory

The persistent object memory $\mathcal{M}_O$ receives streaming video frames $V = [f_1, \ldots, f_n]$, depth maps $D = [d_1, \ldots, d_n]$ and camera poses $P = [p_1, \ldots, p_n]$ as input. It maintains an entry for each perceived object in the 3D scene. Each object entry includes the following fields: an object thumbnail, 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.

3D Object Detection. Given an incoming 2D egocentric frame f_i, depth map d_i, and camera pose p_i, we first employ a 2D open-vocabulary object detection model (YOLO-World [2] in this work) to detect objects, obtaining their categories and 2D bounding boxes from f_i. Based on these detections, we extract **OBJ Feat** and **CTX Feat** by applying CLIP [39] and DINOv2 [36] to the cropped object images and the original frames, respectively. Furthermore, leveraging the depth map d_i and camera pose p_i, we lift the 2D detections to 3D using 2D-to-3D projection [11], obtaining the object's **3D Bbox** in the scene space.

3D Object Re-Identification. After object detection, object re-identification (re-ID) is required to determine whether an object is newly observed or already exists in the memory. We propose an object re-ID algorithm that integrates both 2D visual features and 3D geometry, as illustrated in Algorithm 1. Given the sensory input, objects in $\mathcal{M}_O^{t-1}$ are first

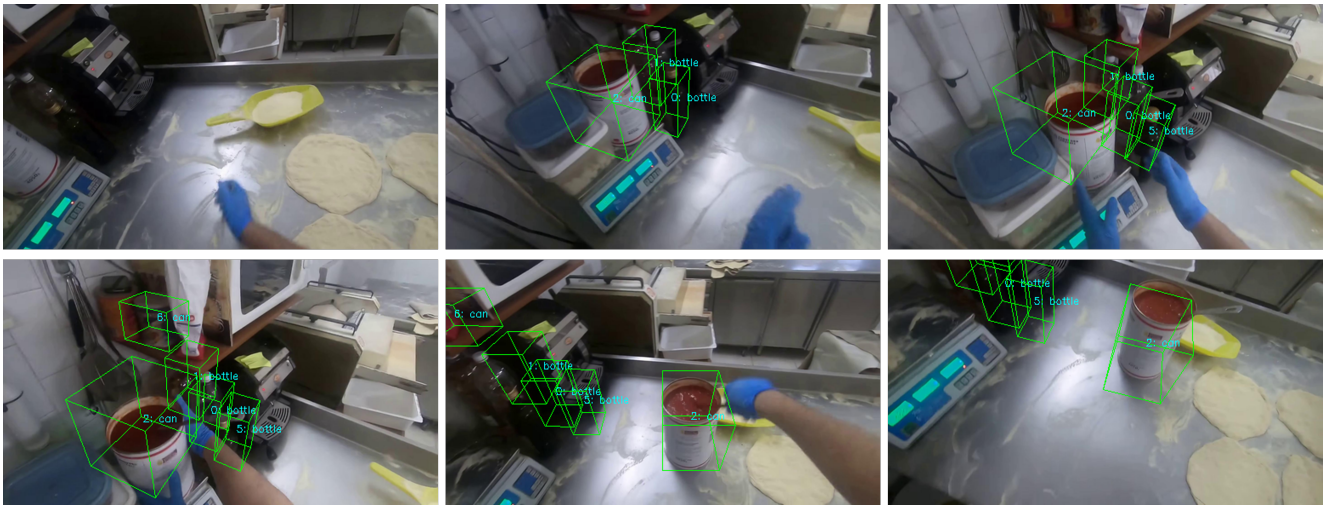


Figure 3. Visualization of the entries in *persistent object memory* $\mathcal{M}_O$. Each 3D bounding box corresponds to an entry in the memory. As the video proceeds, objects (*e.g.* the large canned tomato paste) can be tracked/re-IDed and have their memory entries updated.

Algorithm 1: 3D Object Detection and Re-ID.

Input: current sensory input f_t, d_t, p_t, object memory at the last timestamp $\mathcal{M}_O^{t-1}$.

Output: $\mathcal{M}_O^t$ after the current observations.

$\mathcal{O}_t^{\text{new}} = \text{ObjectDetection}(f_t, d_t, p_t)$ //detection set

$\mathcal{O}_{t-1}^{\text{static}}, \mathcal{O}_{t-1}^{\text{dynamic}} = \text{GeometryCheck}(\mathcal{M}_O^{t-1}, d_t, p_t)$

$\mathcal{M}^{\text{static}} = \text{StaticMatch}(\mathcal{O}_t^{\text{new}}, \mathcal{O}_{t-1}^{\text{static}})$

$\mathcal{O}_t^{\text{static}} = \text{StaticMerge}(\mathcal{M}^{\text{static}}, \mathcal{O}_{t-1}^{\text{static}})$

$\mathcal{O}'^{\text{new}}_t = \mathcal{O}_t^{\text{new}} \setminus \mathcal{M}^{\text{static}}$

$\mathcal{M}^{\text{dynamic}} = \text{DynamicMatch}(\mathcal{O}'^{\text{new}}_t, \mathcal{O}_{t-1}^{\text{dynamic}})$

$\mathcal{O}_t^{\text{static}} = \text{DynamicMerge}(\mathcal{M}^{\text{dynamic}}, \mathcal{O}_{t-1}^{\text{dynamic}})$

$\mathcal{O}''^{\text{new}}_t = \mathcal{O}'^{\text{new}}_t \setminus \mathcal{M}^{\text{dynamic}}$

$\mathcal{M}_O^t = \mathcal{O}_t^{\text{static}} \cup \mathcal{O}_t^{\text{dynamic}} \cup \mathcal{O}''^{\text{new}}_t$

categorized as 'static' or 'dynamic' by comparing their memorized 3D bounding boxes with the current 3D geometry. An object is labeled as 'dynamic' if it has changed position; otherwise, it is considered 'static'. Next, the algorithm attempts to match detected objects with existing entries in the database, leveraging both **OBJ Feat** and **3D Bbox**. For static objects, matching is primarily based on the position of **3D Bbox**, while for dynamic objects, it relies on the similarity of **OBJ Feat** and the size of the **3D Bbox**. Successfully matched objects are merged with their corresponding entries using a moving average on **OBJ Feat**, **CTX Feat**, and **3D Bbox**. Unmatched objects are assigned a new **ID** and added to memory. A visualization of how object entries are created and updated in $\mathcal{M}_O$ can be found in Figure 3. Compared to object re-ID relying solely on 2D features [5], incorporating 3D information leads to improved re-ID performance and enhances the persistence of object memory, making it more responsive to dynamic changes in 3D positions. Additional details can be found in *Appendix*.

Object States and Relationships. Object states (**STATE**) and relationships (**RO**) are crucial for embodied tasks. For instance, knowing whether a fridge door is open or which table the target cup is on directly influences planning and execution. For **RO**, we follow prior work [16, 20, 67] to infer object relationships from their 3D bounding boxes. Currently, we focus on two key relations—"on/uphold" and "in/contain"—as they are most relevant to embodied tasks. The **RO** is updated at each timestamp for all detected objects. The **STATE** field is updated by a VLM when an object is newly detected or interacted with. In this work, **STATE** is limited to "OPEN" or "CLOSED" for articulated receptacles (e.g., fridges, microwaves) and "normal" or "in-hand" for regular receptacles (e.g., tables, beds) and pickable objects. The VLM update procedure is described as follows.

3.2. VLM-based Memory Update

A key challenge in persistent object memory is updating the **STATE** information when actions are performed on objects, especially in the presence of visual occlusion (e.g., hand-can interaction shown in Figure 3). We tackle this by using an action annotator (LaViLa [64]) and a vision-language model (GPT-4o). As shown in Figure 4, when an action (e.g.,"C catches the can") is detected, we first retrieve relevant object entries in $\mathcal{M}_O$ associated with the "can" visible in the current frame (in this case, 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. This visual prompting [1, 59] links the action to the corresponding object entry in memory. Finally, we programmatically update the **STATE** field of the target object entry to "in-hand" after the action "catches the can". The action and the target object ID are then stored in the history buffer. More details are available in *Appendix*.

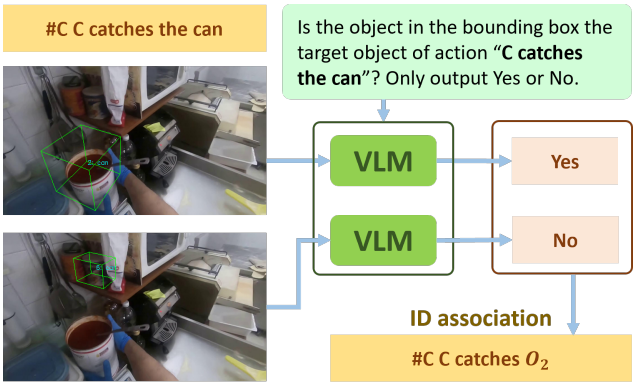


Figure 4. An illustration of our VLM-based memory update method. This approach effectively prompts the VLM to associate an action with relevant object entries in memory through visual prompting, identifying the entries corresponding to the action's target objects.

3.3. Tools and Embodied Action Primitives

Embodied VideoAgent has the following tools for perception: `query_db(·)`, which processes natural language queries to retrieve relevant entries from memory via database queries or similarity-based retrieval using **OBJ Feat**; `temporal_loc(·)`, which retrieves specific video frames corresponding to natural language queries using **CTX Feat** and the history buffer; `spatial_loc(·)`, which determines 3D scene locations based on **CTX Feat** and **3D Bbox** given room queries; and `vqa(·)`, which answers open-ended questions about object thumbnails or video frames.

When equipped with embodied action primitives, the agent can actively explore the scene to complete embodied tasks. These actions are: `chat()` for user interaction; `search(·)` to exhaustively search the scene for specified objects; `goto(·)` for location navigation; and `open(·)`, `close(·)`, `pick(·)`, and `place(·)` for object interactions. The action primitives can be automatically translated into sequences of atomic actions ('move forward', 'turn right', 'tilt up', *etc.*) in the Habitat-Sim simulator [43, 46]. Further implementation details of tools and action primitives are provided in *Appendix*.

4. The Two-Agent Framework

Collecting synthetic data for training 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 framework can automatically generate episodes where a user interacts with an assistant within embodied environments. Drawing inspiration from prior multi-LLM-agent research [8], 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

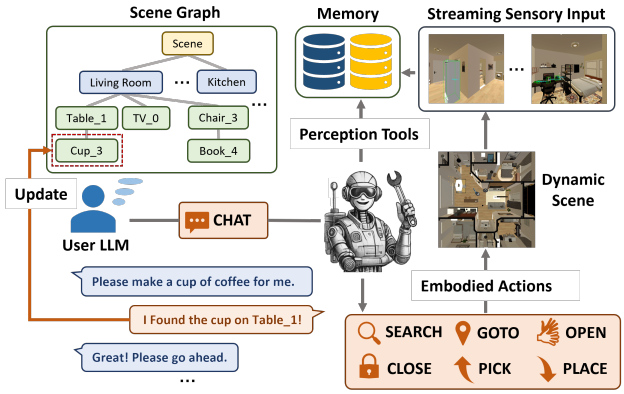


Figure 5. An overview of our synthetic embodied data collection framework. An LLM plays the user role and is prompted to propose engaging tasks based on a partial scene graph and the user's feedback, while the user, effectively a *Embodied VideoAgent*, explores the scene and fulfills the user's requests.

based on its limited scene graph knowledge (achieved by randomly trimming the full scene graph to stimulate curiosity) and the assistant's feedback. Given an interaction episode history, a VLM is prompted to check whether the task is successfully completed and score the quality of the episode based on the chain-of-thought history and the final visual input of *Embodied VideoAgent*. The VLM's scoring standard aligns with that of human evaluators. Detailed prompting strategies are provided in *Appendix*.

5. Capabilities and Analysis

We evaluate *Embodied VideoAgent* from three key perspectives: (1) accurate dynamic scene perception and memory construction, leveraging EgoObjects [66] and Ego4D-VQ3D [9]; (2) embodied question answering on OpenEQA [32] and EnvQA [6]; and (3) open-ended embodied tasks within the two-agent framework, as well as a showcase of robotic manipulation.

5.1. Dynamic Scene Perception

Dynamic scene perception requires constructing an accurate memory focusing on objects within a dynamic setting. We evaluate *Embodied VideoAgent* on Ego4D-VQ3D [9] and EgoObjects [66] to assess its effectiveness.

5.1.1. Ego4D-VQ3D

Ego4D-VQ3D [9] is a benchmark for egocentric video understanding, where a model analyzes first-person video to spatially locate a queried object using its image at a given timestamp. We include three learning-based methods for comparison: EgoLoc [31], Ego4D [9], and Ego4D w/ EgoLoc poses [31]. For zero-shot methods, *Embodied VideoAgent* uses YoloWorld [2] to detect the object in video frames and obtain its 3D position via 2D-3D projection. The query object is retrieved using: (1) category seman-

Table 1. Results of 3D object localization within dynamic scenes on the validation set of Ego4D-VQ3D[9].

Ego4D-VQ3D				
Learning-based Methods				
Method	**Succ%** ↑	**Succ*%** ↑	**L2**↓	**QwP%** ↑
EgoLoc [31]	**80.49**	**98.14**	**1.45**	**82.32**
Ego4D [9]	1.22	30.77	5.98	1.83
Ego4D w/ EgoLoc poses [31]	73.78	91.45	2.05	80.49
Zero-shot Methods				
Method	**Succ%** ↑	**Succ*%** ↑	**L2**↓	**QwP%** ↑
E-VideoAgent(text)	53.05	94.57	2.00	56.10
E-VideoAgent(image)	85.37	92.72	1.86	92.07
E-VideoAgent(text+image)	**86.98**	**95.86**	**1.79**	**93.47**

Table 2. Results on EgoObjects.

EgoObjects (910 objects)	VideoAgent [5]	E-VideoAgent
$\overline{\mathbf{F1}}$	31.76	**53.69**

tic similarity (E-VideoAgent(text)), (2) 2D feature similarity (E-VideoAgent(image)), and (3) a combination of both (E-VideoAgent(text+image)). EgoLoc, Ego4D w/ EgoLoc poses and all variants of *Embodied VideoAgent* are tested on the same pose estimation provided by EgoLoc.

A 2D detector with 3D geometry achieves satisfactory 3D object detection. Table 1 presents the results on the VQ3D validation set. E-VideoAgent(text+image) achieves the highest success rate, outperforming the learning-based EgoLoc by 6.49% on **Succ%**. This highlights the effectiveness of combining an open-vocabulary detector with the 2D-3D lifting approach for 3D object detection. The ablation study on *Embodied VideoAgent* variants indicates that integrating both 2D visual features and detection categories enhances object retrieval accuracy. Notably, 2D visual features contribute more significantly than object category information.

5.1.2. EgoObjects

We evaluate our object re-ID algorithm by incorporating both 2D visual features and 3D geometry, comparing it against VideoAgent [5], which relies solely on 2D visual features. We extract 910 objects across 91 categories from EgoObjects [66], each with a query video frame and a corresponding key frame. Given a query frame, the algorithm must identify its key frame from 10 other key frames within the same category. We compute the $F1$ score for this binary classification and use the average $\overline{F1}$ across all instances for evaluation. For *Embodied VideoAgent*, we employ DUSt3R [50] to construct the 3D geometry of both frames. **3D information contributes to successful re-ID.** As shown in Table 2, leveraging 3D information significantly enhances re-ID performance, with *Embodied VideoAgent* outperforming VideoAgent [5] by 21.93%, demonstrating the necessity of 3D geometry for accurate object re-ID.

Table 3. Results on EM-EQA of OpenEQA [32].

OpenEQA			
Method	**ScanNet**	**HM3D**	**All**
GPT-4 w/ LLaVA-1.5 [32]	45.4	40.0	43.6
GPT-4 w/ CG [32]	37.8	34.0	36.5
GPT-4V (50 frames) [32]	57.4	51.3	55.3
Video-LLaVA [24]	41.5	34.6	39.2
LLaMA-VID [23]	33.4	34.0	33.6
E-VideoAgent(GPT-4o)	**59.4**	**54.9**	**57.9**
Subset (226)			
Method	**ScanNet**	**HM3D**	**All**
GPT-4o w/ LLaVA-1.5 [26]	35.1	36.2	35.6
GPT-4o (8 frames)	43.5	39.8	41.9
Video-LLaVA [24]	32.9	27.8	30.6
LLaMA-VID [23]	31.2	28.0	29.4
VideoAgent [5]	38.9	41.4	40.0
E-VideoAgent(InternVL2)	39.7	43.0	41.2
E-VideoAgent(GPT-4o)	**46.0**	**48.2**	**47.0**

5.2. Embodied Question Answering

Given an embodied episode within a scene, embodied question answering requires the model to answer questions about the scene or the activities. We evaluate *Embodied VideoAgent* on OpenEQA [32] and EnvQA [6].

5.2.1. OpenEQA

OpenEQA [32] is a benchmark for open-ended scene question answering. We evaluate *Embodied VideoAgent* on the full OpenEQA dataset, as well as a subset comprising 126 questions from 10 randomly selected ScanNet scenes and 100 questions from 10 randomly selected HM3D scenes. We compare *Embodied VideoAgent* against three categories of baselines: 1) video understanding models, including GPT-4V, GPT-4o, Video-LLaVA [24], and LLaMA-VID [23]; 2) large language models with scene descriptions, such as GPT-4 with frame captions generated by LLaVA-1.5 [32] and GPT-4 with scene graphs; 3) video understanding agents, specifically VideoAgent [5], which shares the same LLM controller (GPT-4o) as *Embodied VideoAgent*. On the subset, E-VideoAgent(InternVL2) and E-VideoAgent(GPT-4o) utilize InternVL2 and GPT-4o as their VQA tools, respectively. Table 3 shows the results.

Agentic systems outperform strong VLMs. On the full dataset, *Embodied VideoAgent* surpasses GPT-4V with 50 input frames, demonstrating the potential of an LLM-based agent to outperform end-to-end VLMs by leveraging structured memory, multi-step information retrieval, and advanced reasoning capabilities. This advantage is further evident in the OpenEQA subset, where *Embodied VideoAgent* and VideoAgent outperform their respective end-to-end VQA models, GPT-4o and Video-LLaVA. Moreover, the

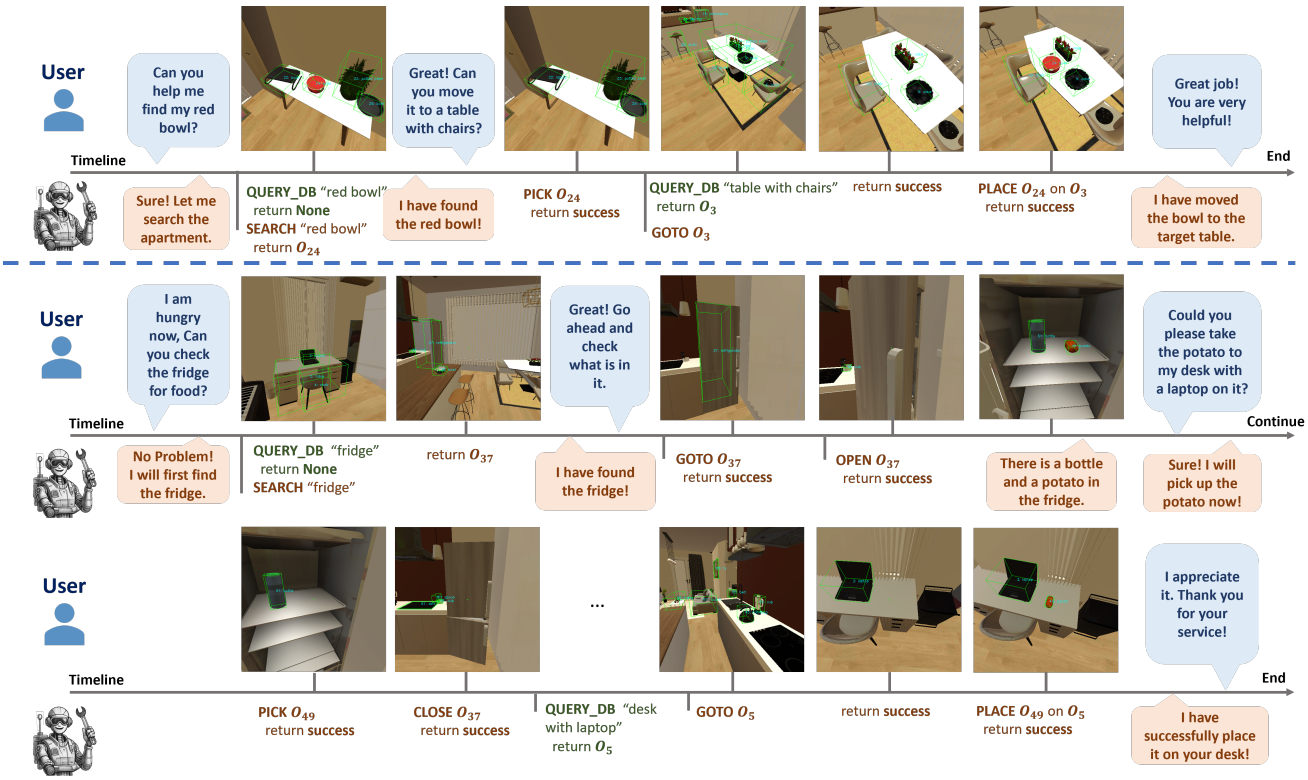


Figure 6. Two episodes of generated embodied user-assistant interaction. The episodes are produced by the framework mentioned in Section 4, where an LLM plays the user and *Embodied VideoAgent* is the assistant. The episodes comprise various embodied problem-solving that require precise memory of the scene objects and tool usage. More example episodes can be found in *Appendix*.

Table 4. Results on the test set of EnvQA [6].

EnvQA				
Method	**Events**	**Orders**	**States**	**All**
Video-LLaVA [24]	10.19	39.00	18.50	22.56
LLaMA-VID [23]	9.98	54.00	5.50	23.16
VideoAgent [5]	5.54	65.50	12.50	27.84
E-VideoAgent	**25.91**	**68.00**	**35.50**	**43.13**
E-VideoAgent w/o object mem.	20.56	55.00	30.50	35.35
E-VideoAgent w/o VLM update	13.78	39.00	29.50	27.42

performance gains of *Embodied VideoAgent* over VideoAgent can be largely attributed to its enhanced object memory with 3D information, which strengthens spatial intelligence. This allows *Embodied VideoAgent* to handle spatial reasoning questions—such as "Where is the picture with the red flowers?"—more effectively.

5.2.2. EnvQA

EnvQA [6] focuses on embodied actions involving object interactions in dynamic scenes. We evaluate *Embodied VideoAgent* alongside strong end-to-end models and video understanding agents on three question types: States (e.g., "Where was the book moved?"), Events (e.g., "What happened after throwing the soap bar?"), and Orders (e.g., "Which happened first: filling the pot with water or using the soap bottle?"). Each category consists of 200 questions. Table 4 presents the results on EnvQA.

Persistent memory and VLM-based update are crucial for action-related tasks. A key factor in accurately answering Events and Orders questions in *Embodied VideoAgent* is the action buffer, which associates each action with its target object. Understanding events in EnvQA is particularly challenging due to the simulated environment, limiting the effectiveness of the action annotator. However, by leveraging persistent object memory to track dynamic objects and using VLM-based updates to identify crucial objects in video frames, *Embodied VideoAgent* can successfully establish action-object associations in dynamic scenes, resulting in improved performance.

5.3. Embodied Tasks

An embodied agent needs to engage in a perception-decision-action loop to achieve a specific goal within an environment. We evaluate *Embodied VideoAgent* within the proposed two-agent framework to assess its task completion capabilities. Additionally, we deploy *Embodied VideoAgent* on a robotic arm for manipulation tasks.

5.3.1. Two-Agent Framework

To evaluate *Embodied VideoAgent*, we randomly generated 15 embodied interaction episodes across 10 different scene

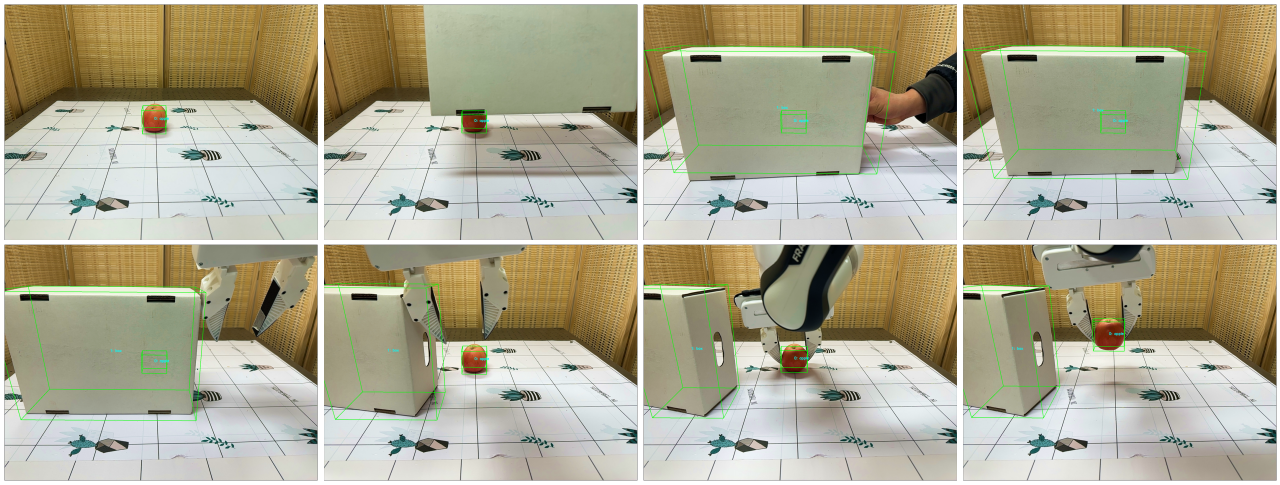


Figure 7. Our persistent object memory enables effective real-world robotic manipulation. Using *Embodied VideoAgent* for perception, the robot is tasked to pick up an apple, which soon becomes occluded by a box. Leveraging its memory, the robot retrieves the apple's position, moves the box aside, and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

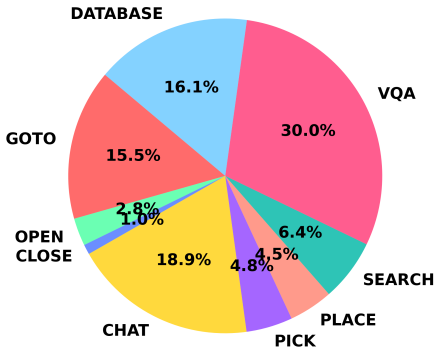


Figure 8. The action distribution of *Embodied VideoAgent*.

Table 5. Statistics in the two-agent framework. T.calls and steps are the average numbers of tool calls and atomic actions (move forward, turn right *etc*.) in an episode.

Two-Agent Framework				
Settings	**Succ**	**Score**	**T. calls**	**Steps**
E-VideoAgent	**0.88**	**4.20**	28.09	1854.34
E-VideoAgent w/o object mem.	0.56	2.73	30.38	2365.13
E-VideoAgent w/o VLM update	0.72	2.88	33.80	1993.83

layouts, resulting in a total of 150 episodes. Figure 6 presents the examples of the episodes and Figure 8 shows the action distribution of *Embodied VideoAgent*. For the ablation study, we compare *Embodied VideoAgent* with two variants: 1) E-VideoAgent w/o object memory, which lacks persistent object memory and performs navigation using visual detection only; 2) E-VideoAgent w/o VLM update, which operates without VLM feedback after executing actions. Table 5 presents the statistics of these variants.

Object memory and visual feedback enhance success rate and efficiency. According to the VLM scorer, *Embodied VideoAgent* achieves an average success rate of 0.88 and an average score of 4.20 out of 5.0, highlighting its effectiveness as an embodied task solver. As shown in Figure 8, `query_db(·)` and `vqa(·)` are the most frequently used tools by *Embodied VideoAgent*. Their importance is further demonstrated in Table 5. When object memory is removed, E-VideoAgent experiences a notable drop in success rate and an increase in execution steps, as it must rely on exhaustive search without prior knowledge of object locations. Similarly, without visual feedback, *Embodied VideoAgent* 's performance and score decline, emphasizing the critical role of VLM-based updates in task success.

5.3.2. Applications in Manipulation

In Figure 7, we demonstrate *Embodied VideoAgent* 's application in robotic perception, where a Franka robot uses it to build persistent memory in a dynamic manipulation task. The robot is instructed to pick up an apple, but it becomes hidden behind a box. Using persistent object memory, the robot recalls the apple's location, moves the box, and successfully completes the task, highlighting the effectiveness of the persistent object memory.

6. Conclusions

We present *Embodied VideoAgent*, a multimodal tool-use agent for dynamic scene understanding. It features a novel persistent object memory and automatic VLM-based updates. Unlike end-to-end multimodal LLMs, *Embodied VideoAgent* integrates egocentric observations with embodied sensory inputs for precise scene comprehension. This approach is robust to dynamic changes, making it ideal for embodied AI tasks. *Embodied VideoAgent* achieves strong results on benchmarks like Ego4D-VQ3D, OpenEQA, and EnvQA. Future work will explore deploying robots in more complex 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: Mastering open-world interaction with visual-temporal context prompting. In *CVPR*, 2025. 4

[2] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. In *CVPR*, 2024. 3, 5, 12

[3] 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. *IJCV*, 2022. 1, 2

[4] 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. 2

[5] Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. Videoagent: A memory-augmented multimodal agent for video understanding. In *ECCV*, 2025. 1, 2, 4, 6, 7, 12

[6] Difei Gao, Ruiping Wang, Ziyi Bai, and Xilin Chen. Env-qa: A video question answering benchmark for comprehensive understanding of dynamic environments. In *ICCV*, 2021. 2, 5, 6, 7

[7] 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. 1, 2

[8] Ran Gong, Qiuyuan Huang, Xiaojian Ma, Hoi Vo, Zane Durante, Yusuke Noda, Zilong Zheng, Song-Chun Zhu, Demetri Terzopoulos, Li Fei-Fei, et al. Mindagent: Emergent gaming interaction. *arXiv preprint arXiv:2309.09971*, 2023. 5

[9] 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 *CVPR*, 2022. 1, 2, 5, 6

[10] Qiao Gu, Ali Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, et al. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. In *ICRA*, 2024. 1, 2

[11] 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

[12] Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *CVPR*, 2023. 1, 2

[13] Tengda Han, Weidi Xie, and Andrew Zisserman. Temporal alignment networks for long-term video. In *CVPR*, 2022. 1

[14] Shengyu Hao, Wenhao Chai, Zhonghan Zhao, Meiqi Sun, Wendi Hu, Jieyang Zhou, Yixian Zhao, Qi Li, Yizhou Wang, Xi Li, et al. Ego3dt: Tracking every 3d object in ego-centric videos. In *ACM MM*, 2024. 2

[15] 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 *NeurIPS*, 2024. 2

[16] 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. In *ICML*, 2024. 1, 4

[17] Krishna Murthy Jatavallabhula, Alihusein Kuwajerwala, Qiao Gu, Mohd Omama, 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, 2

[18] Baoxiong Jia, Yixin Chen, Siyuan Huang, Yixin Zhu, and Song-chun Zhu. Lemma: A multi-view dataset for le arning m ulti-agent m ulti-task a ctivities. In *ECCV*, 2020. 1

[19] Baoxiong Jia, Ting Lei, Song-Chun Zhu, and Siyuan Huang. Egotaskqa: Understanding human tasks in egocentric videos. *NeurIPS*, 2022. 1

[20] Baoxiong Jia, Yixin Chen, Huangyue Yu, Yan Wang, Xuesong Niu, Tengyu Liu, Qing Li, and Siyuan Huang. Sceneverse: Scaling 3d vision-language learning for grounded scene understanding. In *ECCV*, 2025. 4

[21] 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 *CVPR*, 2024. 22

[22] Bruno Korbar, Yongqin Xian, Alessio Tonioni, Andrew Zisserman, and Federico Tombari. Text-conditioned resampler for long form video understanding. In *ECCV*, 2024. 1

[23] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *ECCV*, 2025. 1, 2, 6, 7

[24] Bin Lin, Bin Zhu, Yang Ye, 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, 7

[25] Kevin Qinghong Lin, Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Z Xu, Difei Gao, Rong-Cheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. *NeurIPS*, 2022. 1, 2

[26] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, 2024. 6

[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, Jinmeng 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. 2

[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. 2

[31] Jinjie Mai, Abdullah Hamdi, Silvio Giancola, Chen Zhao, and Bernard Ghanem. Egoloc: Revisiting 3d object localization from egocentric videos with visual queries. In *ICCV*, 2023. 2, 5, 6

[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 *CVPR*, 2024. 2, 5, 6

[33] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *NeurIPS*, 2024. 1, 2

[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. 2

[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. 3, 12

[37] Chiara Plizzari, Shubham Goel, Toby Perrett, Jacob Chalk, Angjoo Kanazawa, and Dima Damen. Spatial cognition from egocentric video: Out of sight, not out of mind. In *International Conference on 3D Vision*, 2025. 2

[38] Shraman Pramanick, Yale Song, Sayan Nag, Kevin Qinghong Lin, Hardik Shah, Mike Zheng Shou, Rama Chellappa, and Pengchuan Zhang. Egovlpv2: Egocentric video-language pre-training with fusion in the backbone. In *ICCV*, 2023. 1, 2

[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 3, 12

[40] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 12

[41] 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

[42] 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 *CVPR*, 2022. 1

[43] 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. 2, 5, 22

[44] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. *NeurIPS*, 2019. 1

[45] Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. In *ICCV*, 2023. 1, 2

[46] Andrew Szot, Alex Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Chaplot, Oleksandr Maksymets, Aaron Gokaslan, Vladimir Vondrus, Sameer Dharur, Franziska Meier, Wojciech Galuba, Angel Chang, Zsolt Kira, Vladlen Koltun, Jitendra Malik, Manolis Savva, and Dhruv Batra. Habitat 2.0: Training home assistants to rearrange their habitat. In *NeurIPS*, 2021. 2, 5

[47] 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

[48] 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, 2

[49] 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, 2

[50] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, 2024. 6, 15

[51] 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. *CoRR*, 2024. 2

[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. 1

[53] 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. *IEEE TPAMI*, 2024.

[54] Zihao Wang, Shaofei Cai, Zhancun Mu, Haowei Lin, Ceyao Zhang, Xuejie Liu, Qing Li, Anji Liu, Xiaojian Shawn Ma, and Yitao Liang. Omnijarvis: Unified vision-language-action tokenization enables open-world instruction following agents. *NeurIPS*, 2024. 1

[55] Olivia Wiles, Joao Carreira, Iain Barr, Andrew Zisserman, and Mateusz Malinowski. Compressed vision for efficient video understanding. In *ACCV*, 2022. 1

[56] 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. 1

[57] Chao-Yuan Wu and Philipp Krahenbuhl. Towards long-form video understanding. In *CVPR*, 2021. 1

[58] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *CVPR*, 2021. 2

[59] 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

[60] Zongxin Yang, Guikun Chen, Xiaodi Li, Wenguan Wang, and Yi Yang. Doraemongpt: Toward understanding dynamic scenes with large language models (exemplified as a video agent). In *ICML*, 2024. 2

[61] 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. 2

[62] 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. 2

[63] 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

[64] Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. Learning video representations from large language models. In *CVPR*, 2023. 4, 14

[65] Yunhan Zhao, Haoyu Ma, Shu Kong, and Charless Fowlkes. Instance tracking in 3d scenes from egocentric videos. In *CVPR*, 2024. 2

[66] Chenchen Zhu, Fanyi Xiao, Andrés Alvarado, Yasmine Babaei, Jiabo Hu, Hichem El-Mohri, Sean Culatana, Roshan Sumbaly, and Zhicheng Yan. Egoobjects: A large-scale egocentric dataset for fine-grained object understanding. In *ICCV*, 2023. 5, 6

[67] 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, 4

[68] 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 *ECCV*, 2024. 1

[69] Ziyu Zhu, Xilin Wang, Yixuan Li, Zhuofan Zhang, Xiaojian Ma, Yixin Chen, Baoxiong Jia, Wei Liang, Qian Yu, Zhidong Deng, Siyuan Huang, and Qing Li. Move to understand a 3d scene: Bridging visual grounding and exploration for efficient and versatile embodied navigation. *ICCV*, 2025. 1