



EgoPlan-Bench: Benchmarking Multimodal Large Language Models for Human-Level Planning

Yi Chen¹ · Yuying Ge² · Yixiao Ge^{2,3} · Mingyu Ding⁴ · Bohao Li¹ · Rui Wang¹ · Ruifeng Xu⁵ · Ying Shan^{1,3} · Xihui Liu² 

Received: 30 September 2024 / Accepted: 23 September 2025
© The Author(s) 2026

Abstract

The pursuit of artificial general intelligence (AGI) has been accelerated by Multimodal Large Language Models (MLLMs), which exhibit superior reasoning, generalization capabilities, and proficiency in processing multimodal inputs. A crucial milestone in the evolution of AGI is the attainment of human-level planning, a fundamental ability for making informed decisions in complex environments, and solving a wide range of real-world problems. Despite the impressive advancements in MLLMs, a question remains: **How far are current MLLMs from achieving human-level planning?** To shed light on this question, we introduce EgoPlan-Bench, a comprehensive benchmark to evaluate the planning abilities of MLLMs in real-world scenarios from an egocentric perspective, mirroring human perception. EgoPlan-Bench emphasizes the evaluation of planning capabilities of MLLMs, featuring realistic tasks, diverse action plans, and intricate visual observations. Our rigorous evaluation of a wide range of MLLMs reveals that EgoPlan-Bench poses significant challenges, highlighting a substantial scope for improvement in MLLMs to achieve human-level task planning. To facilitate this advancement, we further present EgoPlan-IT, a specialized instruction-tuning dataset that effectively enhances model performance on EgoPlan-Bench. We have made all the codes, data, and a maintained benchmark leaderboard available at https://chenyi99.github.io/ego_plan/ to advance future research.

Keywords Multimodal large language model benchmark · Human-level planning · Egocentric video · Real-world scenario

1 Introduction

Multimodal Large Language Models (MLLMs) combine the exceptional reasoning and generalization capabilities of Large Language Models (Brown et al., 2020; Ouyang et

al., 2022; Touvron et al., 2023; Chiang et al., 2023) with the capacity to understand multimodal inputs. This integration has opened new possibilities for achieving the ultimate goal of artificial general intelligence (AGI), which aims to match or surpass human performance in most tasks (y Arcas & Norvig, 2023; Bubeck et al., 2023; Morris et al., 2023). A crucial milestone in the evolution of AGI is the achievement of human-level planning capabilities, as task planning constitutes an essential aspect of everyday human activities, enabling informed decision-making in complex environments and the resolution of numerous real-world problems.

Task planning involves predicting the subsequent action plan based on the task goal, real-time progress, and current observations, as illustrated in Figure 2. This capability is indispensable for a model to be effectively applied in the real world for addressing complex problems, in the same manner as humans. It should be regarded as one of the fundamental criteria for determining whether MLLMs have genuinely attained human-level intelligence. However, achieving human-level planning for MLLMs is far from triv-

Communicated by Yin Li.

✉ Yuying Ge
yuyingge@tencent.com

✉ Xihui Liu
xihuiliu@eee.hku.hk

¹ The University of Hong Kong, HongKong SAR, China

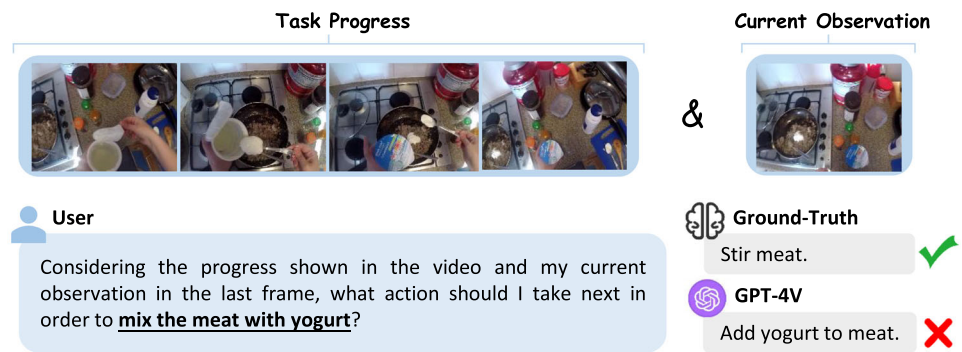
² AI Lab, Tencent, Shenzhen 518052, Guangdong Province, China

³ ARC Lab, Tencent PCG, Shenzhen 518052, Guangdong Province, China

⁴ University of California, Berkeley 94720, CA, USA

⁵ Peng Cheng Laboratory, Shenzhen 518052, Guangdong Province, China

Fig. 1 EgoPlan-Bench presents significant challenges for human-level planning given real-time task progress and real-world visual observations (even GPT-4V makes mistakes)



ial. For example, as shown in Figure 1, to determine the next action plan for the goal “mixing the meat with yogurt”, the model must track several state changes of the yogurt container and spoon, identify the white yogurt paste on the meat, and recognize that further stirring is necessary. Even advanced GPT-4V could make wrong decisions.

This raises the question: “**How far are current MLLMs from achieving human-level planning?**” To investigate this, we introduce a comprehensive and rigorous benchmark called **EgoPlan-Bench** to evaluate the planning abilities of MLLMs in real-world scenarios from an egocentric perspective, which emulates human perception of the world. Unlike existing benchmarks (Fan, 2019; Jia et al., 2022; Li et al., 2023) that mainly focus on the **comprehension** ability of models, EgoPlan-Bench takes a step forward by systematically studying the **planning** capability of MLLMs. As shown in Figure 2, to answer questions in EgoPlan-Bench correctly, the model must be capable of understanding open-form task goals, tracking long-horizon task progress, perceiving real-world environments from an egocentric view, and utilizing inherent world knowledge to reason about the next action plan in an end-to-end manner.

To construct EgoPlan-Bench, we leverage large-scale egocentric videos (Damen et al., 2022; Grauman et al., 2022) which authentically reflect daily human activities from the first-person perspective. We design a semi-automatic data construction pipeline (see Figure 3), and all the evaluation data has undergone strict manual verification. Our benchmark exhibits three main characteristics: **1) Realism of Tasks:** The tasks are extrapolated from authentic real-world videos, offering a closer reflection of daily human needs and showcasing greater variety than artificially designed tasks. **2) Diversity of Action Plans:** The benchmark involves a diverse set of action plans, requiring interaction with hundreds of different objects and extending beyond basic manipulation skills such as picking and placing items. **3) Intricacy of Visual Observations:** The visual observations come across various real-world scenes, where objects vary in appearance, state, and placement. Besides, the visual inputs can span extensive periods, making it challenging for moni-

toring task progress and detecting critical changes in object states.

We evaluate a total of 28 MLLMs on our benchmark. The results indicate that our benchmark poses significant challenges for existing MLLMs, and there is still a long way to go before these models evolve into human-level task planners. We further construct an instruction-tuning dataset **EgoPlan-IT**, specialized for enhancing human-level planning. The model fine-tuned on this dataset demonstrates robust performance gains on the proposed benchmark, verifying the effectiveness of our data.

In summary, our main contributions include:

- We introduce an MLLM benchmark EgoPlan-Bench for human-level planning in real-world scenarios, which features realistic tasks, diverse actions, and intricate visual observations.
- We evaluate a wide range of MLLMs and find that our benchmark poses significant challenges for them. We further provide a detailed analysis of the possible reasons.
- We construct an instruction-tuning dataset EgoPlan-IT specifically for human-level planning, which effectively improves model performance on EgoPlan-Bench.
- We have released all the codes and data, and are consistently maintaining a leaderboard to advance future research.

2 Related Work

Multimodal Large Language Models (MLLMs). Recent advancements in Large Language Models (LLMs) (Ouyang et al., 2022; Touvron et al., 2023; Chiang et al., 2023; Chung et al., 2022) have led to significant achievements in language understanding and generation. As a result, there is a growing interest in developing Multimodal Large Language Models (MLLMs) that combine LLMs’ impressive language capabilities with multi-modal perception abilities (Zhang et al., 2023; Ye et al., 2023; Li et al., 2023; Gong et al., 2023; Awadalla et al., 2023; Gao et al., 2023; Peng et al., 2023, ?;

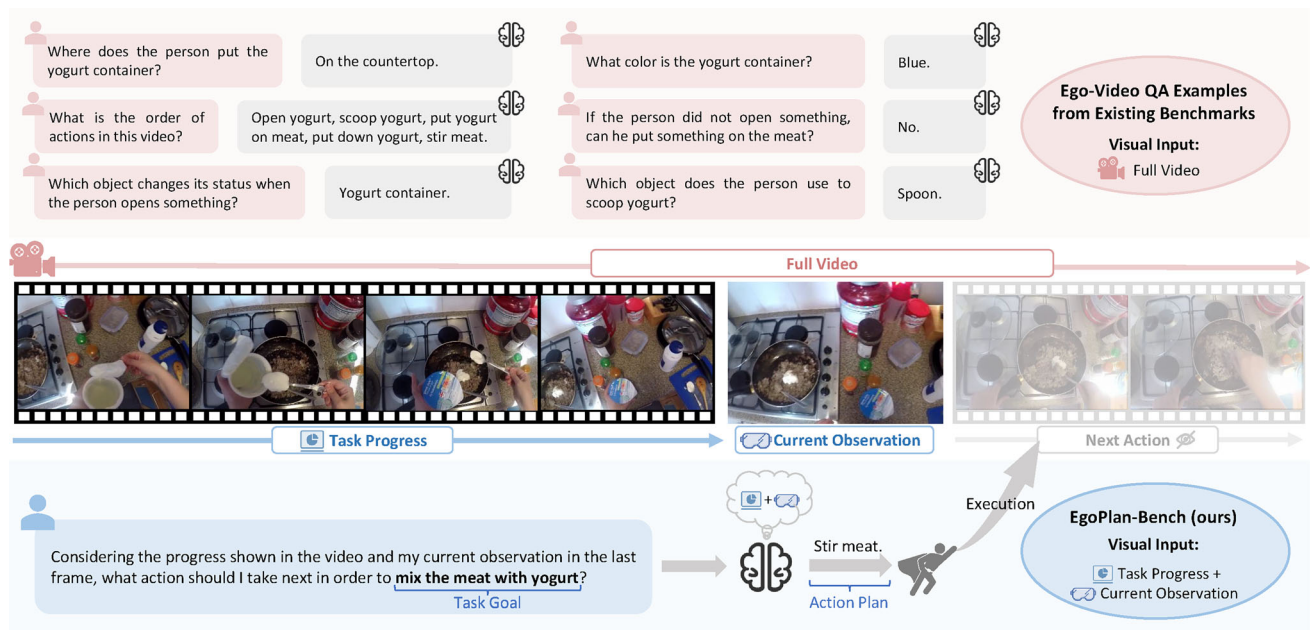


Fig. 2 Our EgoPlan-Bench evaluates **Planning**, where a model predicts the next feasible action plan by taking a video showing task progress, current visual observation, and open-form task goal as inputs like humans. In contrast, the egocentric-video-based QA examples from

existing benchmarks mainly evaluate **Comprehension**, where a model answers questions based on the spatial and temporal understanding of the entire video

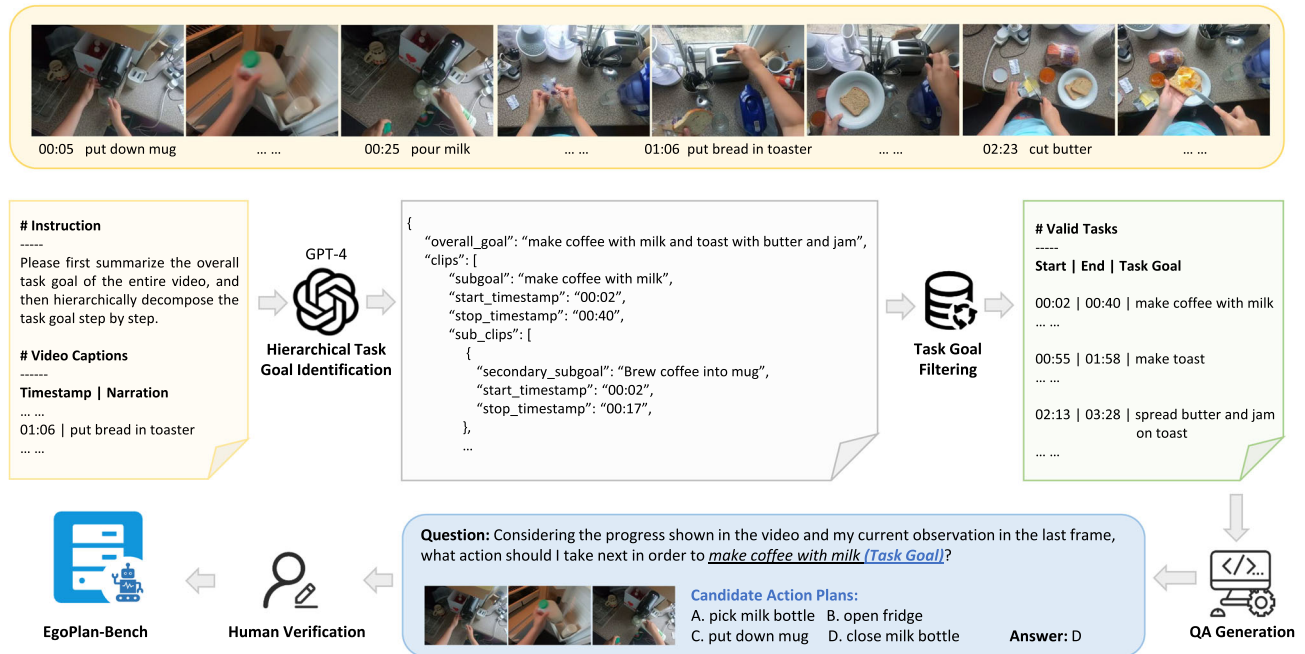


Fig. 3 Overview of the construction pipeline for EgoPlan-Bench based on existing untrimmed egocentric videos with detailed action narrations. (1) We first leverage GPT-4 to identify task goals through hierarchical reasoning. (2) We then filter task goals based on the requisite number of actions. (3) The questions are designed in the form of multiple-choice,

where the questions are automatically generated based on task goals, and the options are derived from different actions under the same task goal. (4) We employ human annotators to verify each question to ensure the benchmark quality

Bai et al., 2023; Zhang et al., 2023; Wang et al., 2023). Typical research involves integrating representations from pretrained visual encoders into LLMs' input embedding space or the intermediate layers through pretraining on image-text interleaved datasets (Li et al., 2023; Dai et al., 2023; Zhu et al., 2023; Liu et al., 2023a, b; Ye et al., 2023; Lu et al., 2024; Young et al., 2024). Recent studies extend this approach to incorporate video inputs, enhancing LLMs' video comprehension capabilities (Zhang et al., 2023; Li et al., 2023; Maaz et al., 2023). Some works even incorporate MLLMs with visual generation abilities (Ge et al., 2023a, b, 2024). In addition, closed-source models such as GPT4V (xxxx, 2023) and Gemini-vision-Pro (Team et al., 2023) have demonstrated leading performance among existing MLLMs.

MLLM Benchmarks. The rapid advancement of MLLMs has brought us closer to achieving AGI, with their evaluation playing a crucial role in their evolution. Numerous benchmarks, such as MME (Fu et al., 2023), LAMM (Yin et al., 2024), Seed-Bench (Li et al., 2023), MMBench (Liu et al., 2023), MM-VET (Yu et al., 2023), MMMU (Yue et al., 2023), LVLm-eHub (Xu et al., 2023), and MMT-Bench (Ying et al., 2024), have been proposed to assess MLLM performance across various dimensions. However, these general-purpose benchmarks allocate only a small proportion to egocentric vision, focusing mainly on comprehension capabilities. Specialized benchmarks such as EgoThink (Cheng et al., 2023) and PCA-Eval (Chen et al., 2023) address egocentric thinking and embodied decision-making, respectively. However, they are constrained by the limited number of handcrafted questions for planning and their reliance on single-image visual observation. Additionally, PCA-Eval's visual inputs are largely derived from simulations, which differ significantly from real-world observations. In this paper, we introduce a comprehensive MLLM benchmark emphasizing human-level planning, incorporating open-form language goals, task progress videos, and real-world egocentric observations. Our dataset consists of thousands of evaluation instances, training instances, and held-out test instances, and it is used to assess a wide range of MLLMs.

Anticipation and Planning in Videos. Human-level planning (HP) is closely related to future anticipation and planning in videos, which aim to predict future states or actions based on observed visual context. Future anticipation methods forecast upcoming conditions or actions from video clips (Sener & Yao, 2019; Damen et al., 2022; Grauman et al., 2022; Abu Farha et al., 2021; Furnari et al., 2017; Liu et al., 2022; Lotter et al., 2016; Villegas et al., 2017; Mendonca et al., 2023), but often do not explicitly incorporate goal-directed planning. In contrast, goal-oriented action prediction—anticipating actions conditioned on specific goals—is more applicable to real-world

scenarios requiring purposeful behavior. Procedural activity understanding and planning have been extensively explored in recent literature. For instance, Lin et al. (2022) propose learning procedural activity recognition with distant supervision, enabling models to comprehend complex multi-step tasks from weakly labeled data. Similarly, Zhong et al. (2023) develop procedure-aware video representations by leveraging instructional videos and their narrations. These studies underscore the importance of modeling procedural knowledge and temporal dependencies for effective video-based planning. The procedure planning task, introduced by Chang et al. (2020), predicts intermediate actions given start and end images. Notable works in this line include Bi et al. (2021), who address procedure planning through contextual modeling and model-based policy learning; Sun et al. (2022), who learn structured and plannable state and action spaces directly from unstructured videos for visually grounded planning; and Zhao et al. (2022), who introduce a probabilistic procedure planning approach under weak supervision. However, specifying goals via images is less natural than using language instructions, and the absence of a terminating visual state limits the applicability of such procedure planning methods in real-world assistance scenarios. VPA (Patel et al., 2023) advances this line of work by incorporating language-specified goals and progress videos to predict task-advancing actions. Nevertheless, VPA primarily relies on instructional YouTube videos (Zhukov et al., 2019; Tang et al., 2019), which often feature third-person perspectives and scripted content, thereby limiting their alignment with authentic egocentric human experiences. Moreover, VPA has not been extensively benchmarked with Multimodal Large Language Models (MLLMs). Egocentric Action Frame Generation (Lai et al., 2023) is another related task focusing on egocentric vision but mainly addressing short-term visual synthesis for atomic actions, whereas HP emphasizes long-horizon, goal-directed planning.

Egocentric Videos. Egocentric videos provide a distinctive perspective for active engagement with the real world. Over the years, numerous egocentric video datasets have been developed (Pirsiavash & Ramanan, 2012; Li et al., 2015; Sigurdsson et al., 2018). Large-scale daily lifelog datasets such as Epic-Kitchens (Damen et al., 2022) and Ego4D (Grauman et al., 2022) offer fine-grained information on human-object interactions, making them valuable resources for studying human-level planning. Existing egocentric video QA benchmarks primarily assess models' understanding of activities and objects in a given egocentric video demonstration. These questions range from visual concept recognition (Fan, 2019) to spatial, temporal, and causal relationship reasoning (Jia et al., 2022). However, none of these benchmarks address the specific needs of human-level planning, where a more practical question and answer would involve a task goal instruction

and a prediction of the next appropriate action that has not yet occurred but is helpful for achieving the given task goal. Song et al. (2024) recently introduces an extended set of goal-oriented activity labels for Ego4D. However, their focus remains on the understanding abilities of models, including goal/step localization and online goal/step detection, rather than addressing the aforementioned gap in human-level planning.

3 EgoPlan-Bench

As illustrated in Figure 2, given a language instruction specifying the task goal, Human-level Planning aims to decide the next appropriate action plan based on the visual input. The visual input includes a video showing a sequence of historical frames capturing the task progress, with the last frame indicating the current egocentric observation. The proposed EgoPlan-Bench contains 5K high-quality human-verified multiple-choice questions derived from egocentric videos (Damen et al., 2022; Grauman et al., 2022). In Sec. 3.1, we first introduce the video sources we used to build the benchmark. We illustrate how to automatically extract task goals from these videos in Sec. 3.2 and structure questions based on the extracted task goals in Sec. 3.3. We then report the detailed data statistics in Sec. 3.4 and describe our evaluation strategy in Sec. 3.5. The entire benchmark construction pipeline is demonstrated in Figure 3.

3.1 Egocentric Video Sources

Unscripted, egocentric videos encompass a wide range of daily human tasks and diverse actions, making them an ideal and challenging medium for evaluating human-level planning. We select two large-scale and well-annotated datasets, Epic-Kitchens (Damen et al., 2022) and Ego4D (Grauman et al., 2022), as our video sources. In Epic-Kitchens, each video is annotated with dense narrations and start/end timestamps of actions. The narrations are expressed as brief verb-object phrases. Ego4D offers similar annotations, except that it only provides a single timestamp for each action and the narrations are expressed as brief sentences like “*C does something*”. We convert Ego4D narrations to verb-object phrases as Epic-Kitchens with GPT-4 (OpenAI, 2023) and derive the start and end timestamps for each action segment following EgoVLP (Lin et al., 2022).

3.2 Task Goal Extraction

Hierarchical Task Goal Identification. Egocentric videos, which capture tasks ranging from simple (involving fewer than three actions) to complex (entailing a long sequence of actions), present challenges in identifying task goals due to

their varying lengths. To address this, we introduce a hierarchical reasoning strategy to identify task goals. As illustrated in Figure 3, given the dense narrations of actions, we utilize GPT-4 to decompose video content into three levels: overall goal $\rightarrow$ subgoal $\rightarrow$ secondary subgoal. For instance, the goal to “*make coffee with milk and toast with butter and jam*” is divided into the subgoals of “*making coffee with milk*” and “*making toast with butter and jam*,” with the former further broken down into secondary subgoals like “*brew coffee into mug*” and “*merge coffee and milk*.” This hierarchical approach allows GPT-4 to efficiently summarize tasks of varying complexity by organizing them into a structured framework, similar to chain-of-thought reasoning. Our findings suggest that prompting GPT-4 to generate all task goals at once may lead to difficulties due to the extensive information presented in longer videos. The detailed prompt can be found in Figure 4.

Task Goal Filtering. Given the variance in video lengths, the task hierarchy across different videos might not align consistently. For example, a short video with only a few actions likely represents a single, simple task, so GPT-4 may only identify one unique goal without further subdivision. In contrast, a long video with many actions often involves a more complex task hierarchy, where the overall goal is broad and subgoals are more specific. We assume that tasks of similar complexity and granularity typically involve a comparable number of actions. Therefore, we aggregate the overall goals, subgoals, and secondary subgoals extracted from various videos under the unified term “task goals”, and refine this set by filtering based on the requisite number of actions. Only task goals that require between 4 and 20 actions are retained. Focusing on task goals within this moderate action range helps standardize the complexity of tasks, making comparisons more consistent and meaningful.

3.3 Multiple-Choice Question Answering

Following the mainstream MLLM benchmarks (Li et al., 2023; Liu et al., 2023; Yue et al., 2023; Ying et al., 2024), we evaluate human-level planning with multiple-choice questions, which facilitates the convenient computation of accuracy as an objective metric. Furthermore, such a question structure allows us to carefully design negative choices to control the quality and difficulty of the EgoPlan benchmark.

Automatic QA Generation. Using the filtered task goals, we automatically generate questions through rule-based templates. As illustrated in the blue box of Figure 3, one such template is: “*Considering the progress shown in the video and my current observation in the last frame, what action should I take next in order to [TASK GOAL]?*” We create questions by substituting the placeholder “[TASK GOAL]” with various specific task goals. For each task goal, multiple

The following provides timestamped narrations about actions from an ego-centric video. Your tasks are:

1. Carefully read the provided narrations of this ego-centric video, and connect the narrations by timestamps. While reading the narrations, please pay attention to inferring the temporal relationships of actions, pre- and post- conditions of actions, object affordance, object states, and possible changes of functional zones that support different activities, etc.
2. Recognize the overall goal of the actions in this ego-centric video.
3. Temporally segment this video into clips. Actions in the same clip complete the same subgoal that is helpful for achieving the overall goal.
4. Further segment each clip into more fine-grained sub-clips. Actions in the same sub-clip are likely to take place in a compact space of the same functional zone without moving around, and focus on a secondary subgoal that is helpful for achieving the corresponding parent-level subgoal.

All the related timestamps in the output should be copied from the given timestamped narrations; the descriptions of the overall goal, subgoals, and secondary subgoals should be detailed with related essential objects; and the output must follow the JSON format:

```
{"overall_goal": ... , "clips": [{"subgoal": ..., "start_timestamp": ..., "stop_timestamp": ...,
"sub_clips": [{"secondary_subgoal": ..., "start_timestamp": ..., "stop_timestamp": ...}, ...]}, ...]}
```

Begin!

Timestamped narrations are:

start_timestamp	stop_timestamp	narration
...		
00:00:08.70	00:00:09.56	open fridge
00:00:09.64	00:00:10.64	pick milk
00:00:20.73	00:00:21.63	close fridge
... ..		

Fig. 4 The detailed prompt for task goal identification with GPT-4

question-answer (QA) pairs can be generated by varying the task progress and current observations. For instance, if a task goal involves N distinct actions, we can produce N corresponding QA pairs, each with a different action serving as the ground-truth answer.

Visual Input Alignment. We align the visual input with the ground-truth answer for each QA pair. The visual input contains the historical observations since the whole task begins and right before the action described by the ground-truth answer occurs. The last frame of the visual input indicates the

current observation for planning the next action. To prevent the models from cheating with the clues from hand-object interaction (e.g., a hand touching a tap implies turning it on or off), we set the end timestamp of the visual input as 0.5 seconds preceding the actual next-action segment.

Negative Choice Selection. For each QA pair, we randomly sample three distinct actions with the same task goal, excluding the ground-truth action, as negative choices. This challenging setting demands a deep understanding of the environment state from dynamic visual input and world knowledge, such as action order dependency, rather than just the semantic meanings of task goals and actions, to discern the correct action plan. All candidate options appear reasonable for the given task goal if the visual modality is ignored. To maintain consistent granularity among options and ensure a unique correct answer, each multiple-choice question has been verified by human annotators.

Human Verification. We employ 10 expert annotators with at least a bachelor's degree to meticulously verify the accuracy of each auto-generated multiple-choice question. Initially, annotators must pass a quality assurance phase by pre-annotating 50 samples according to comprehensive guidelines, which are then reviewed by two senior annotators. During the main annotation phase, annotator pairs independently select the most suitable answers for each question based on the task goal and visual cues. They also flag questions with unclear visuals or ambiguous options for removal. The detailed annotation guidelines are provided in Table 1. We only retain questions with unanimous annotator agreement and alignment with the ground-truth answer, resulting in 5K high-quality evaluation instances closely correlated with visual observations.

3.4 Data Statistics

The evaluation data of EgoPlan-Bench, detailed in Table 2, includes 4,939 multiple-choice questions covering 3,296 task goals. Figure 6 shows the distribution of task goals. Overall, there are 3,185 action plans (verb-object phrases) from the questions' candidate options, involving 234 distinct root verbs and 558 unique objects. The most common action plans are illustrated in Figure 5, demonstrating the benchmark's variety of action plans. Additionally, the visual observations of these questions originate from egocentric videos taken across 419 different scenes. This diversity in task goals, candidate actions, and visual scenes underscores the challenges presented by our benchmark. We divide the evaluation data into two subsets: EgoPlan-Val for validation and EgoPlan-Test for testing. To ensure a fairer public comparison of model performance and to enable future opportunities for hosting related challenges, the ground-truth answers of EgoPlan-Test are kept private.

Table 1 Guidelines for Human Verification

Annotation Task Overview

Each annotation example consists of four components:

1. A Question describing the Task Goal, which may involve baking, doing laundry, preparing ingredients, tidying up the room, etc.
2. An Egocentric Video Clip representing the Task Progress, which captures a series of completed activities related to the current task.
3. A Current Observation image that illustrates the environmental state from the egocentric view before performing the next action.
4. Four Candidate Choices outlining potential action plans.

Your task is to infer the task progress by considering the Task Goal from the Question and the Task Progress video. Then, select the most reasonable next action plan from the Candidate Choices that effectively advances the Task Goal and is feasible within the Current Observation's environment.

Annotation Process

1. Begin by reading and understanding the Question, focusing on the Task Goal it addresses.
2. Carefully watch the Task Progress video, identifying state changes of key objects related to the Task Goal, and infer the Task Progress.
3. Examine the Current Observation image, identifying potential target objects for operation within the scene.
4. Choose the most reasonable next action from the Candidate Choices (A~D) that advances the task and can be executed within the Current Observation. This action should:
 - i) Align with the Task Goal specified in the Question, rather than completing unrelated tasks.
 - ii) Involve target objects that are clearly visible within the Current Observation.
 - iii) Prioritize actions that should occur earliest when multiple actions meet the conditions, based on the analysis of the sequence of actions.
 - iv) Not be the action that is already taking place in the Current Observation.

If none of the Candidate Choices (A~D) meet the above criteria, or if the task goal description is ambiguous, or if it is impossible to accurately assess the task progress, or if more than one reasonable option exists, select "E (Skip)".

3.5 Evaluation Strategy

Inspired by Brohan et al. (2023), Dai et al. (2023), Brown et al. (2020), Lin et al. (2021), Li et al. (2023), we use the closed-set answer ranking strategy based on language completion distribution as the major evaluation approach, rather than directly matching a model's free-form generation results with candidate options.

Given a candidate action plan $\mathbf{a} = (a_1, a_2, \dots, a_n)$, which is a sequence of tokens, we compute its log-likelihood under the MLLM by factorizing it as the average of the conditional log-likelihoods of each token given the preceding tokens, the visual input $\mathbf{v}$, and the user instruction $\mathbf{l}$:

Table 3 Performance of MLLMs on EgoPlan-Val, evaluated by closed-set answer ranking

Model	Visual Encoder	LLM	Acc%
Random	-	-	24.32
BLIP-2 (Li et al., 2023)	EVA-CLIP ViT-g/14 (Fang et al., 2023)	Flan-T5-XL	26.71
InstructBLIP (Dai et al., 2023)	EVA-CLIP ViT-g/14 (Fang et al., 2023)	Flan-T5-XL	28.09
InstructBLIP Vicuna (Dai et al., 2023)	EVA-CLIP ViT-g/14 (Fang et al., 2023)	Vicuna-7B	26.53
LLaVA (Liu et al., 2023a)	CLIP ViT-L/14 (Radford et al., 2021)	LLaMA-7B	27.00
MiniGPT-4 (Zhu et al., 2023)	EVA-CLIP ViT-g/14 (Fang et al., 2023)	Vicuna-7B	28.11
VPGLTrans (Zhang et al., 2023)	EVA-CLIP ViT-g/14 (Fang et al., 2023)	LLaMA-7B	27.38
MultiModal-GPT (Gong et al., 2023)	CLIP ViT-L/14 (Radford et al., 2021)	Vicuna-7B	27.81
Otter (Li et al., 2023)	CLIP ViT-L/14 (Radford et al., 2021)	LLaMA-7B	28.08
OpenFlamingo (Awadalla et al., 2023)	CLIP ViT-L/14 (Radford et al., 2021)	LLaMA-7B	27.67
LLaMA-Adapter V2 (Gao et al., 2023)	CLIP ViT-L/14 (Radford et al., 2021)	LLaMA-7B	27.81
GVT (Wang et al., 2023)	Good Visual Tokenizer (Wang et al., 2023) distilled from EVA-CLIP ViT-L/14 (Fang et al., 2023)	Vicuna-7B	27.87
mPLUG-Owl (Ye et al., 2023)	CLIP ViT-L/14 (Radford et al., 2021)	LLaMA-7B	27.63
Kosmos-2 (Peng et al., 2023)	CLIP ViT-L/14 (Radford et al., 2021)	Decoder only 1.3B	26.97
Qwen-VL-Chat (Bai et al., 2023)	Openclip ViT-bigG (Ilharco et al., 2021)	Qwen-7B	27.69
LLaVA-1.5 (Liu et al., 2023b)	CLIP ViT-L/14 (Radford et al., 2021)	Vicuna-7B	27.81
VideoChat (Li et al., 2023)	EVA-CLIP ViT-g/14 (Fang et al., 2023)	Vicuna-7B	27.51
Video-ChatGPT (Maaz et al., 2023)	CLIP ViT-L/14 (Radford et al., 2021)	LLaMA-7B	27.33
Valley (Luo et al., 2023)	CLIP ViT-L/14 (Radford et al., 2021)	LLaMA-13B	27.27
Video-LLaMA (Zhang et al., 2023)	EVA-CLIP ViT-g/14 (Fang et al., 2023)	LLaMA2-Chat-7B	28.58
SEED-LLaMA (Ge et al., 2023b)	EVA-CLIP ViT-g/14 (Fang et al., 2023)	LLaMA2-Chat-13B	29.93
CogVLM (Wang et al., 2023)	EVA2-CLIP-E (Fang et al., 2023)	Vicuna-7B	27.48
DeepSeek-VL-Chat (Lu et al., 2024)	(SigLIP-L (Zhai et al., 2023) and ViTDet (Li et al., 2022))	DeepSeek-LLM-7B	27.57
mPLUG-Owl-2 (Ye et al., 2023)	CLIP ViT-L/14 (Radford et al., 2021)	LLaMA2-7B	27.84
Yi-VL (Young et al., 2024)	Openclip ViT-H/14 (Ilharco et al., 2021)	Yi-6B	28.67
SEED-X (Ge et al., 2024)	Openclip ViT-bigG (Ilharco et al., 2021)	LLaMA2-Chat-13B	31.07
XComposer (Zhang et al., 2023)	EVA-CLIP ViT-g/14 (Fang et al., 2023)	InternLM-7B	37.17

2023; Li et al., 2024, 2023), for some MLLMs that were not trained to take in multiple images, we uniformly sample 8 key frames for each video and concatenate their visual features before feeding them into the language model backbones of MLLMs.

4.2 Performance Analysis

In this paper, we mainly evaluate various MLLMs on the EgoPlan-Val set, with results on the held-out EgoPlan-Test set available in our online leaderboard¹. Table 3 and Table 4 show that most models perform only slightly better than random guessing, with the best model (GPT-4V) achieving just 37.98% accuracy. This indicates that current MLLMs are still far from being sophisticated in human-level planning.

¹ https://huggingface.co/spaces/ChenYi99/EgoPlan-Bench_Leaderboard

Case Study with GPT-4V Since GPT-4V achieves the best performance among all the evaluated MLLMs, we use it for a qualitative case study to delve deeper into why current MLLMs may underperform on our benchmark. As depicted in Figure 7, we require GPT-4V to explicitly summarize the task progress and describe the current observation before planning the next action.

Our observations indicate that while GPT-4V successfully identifies obvious actions, like “*removing the lid of yogurt container*” or “*measuring yogurt with a spoon*,” it might overlook more subtle state changes that are critical to the task’s objective. For instance, GPT-4V fails to recognize that yogurt has been added to the meat, as indicated by the white paste visible on the meat post-measurement. Furthermore, even when the yogurt has been placed on the countertop in the final frame, GPT-4V erroneously maintains that the current observation shows “*a hand holding a spoon containing yogurt*.” This misinterpretation of the visual input leads to

Table 4 Performance of MLLMs on EgoPlan-Val, evaluated by explicit candidate option selection

Model	Visual Encoder	LLM	Acc%
Random	-	-	24.32
BLIP-2 (Li et al., 2023)	EVA-CLIP ViT-g/14 (Fang et al., 2023)	Flan-T5-XL	26.43
InstructBLIP (Dai et al., 2023)	EVA-CLIP ViT-g/14 (Fang et al., 2023)	Flan-T5-XL	26.68
InstructBLIP Vicuna (Dai et al., 2023)	EVA-CLIP ViT-g/14 (Fang et al., 2023)	Vicuna-7B	24.72
Valley (Luo et al., 2023)	CLIP ViT-L/14 (Radford et al., 2021)	LLaMA-13B	26.80
Qwen-VL-Chat (Bai et al., 2023)	Openclip ViT-bigG (Ilharco et al., 2021)	Qwen-7B	30.50
LLaVA-1.5 (Liu et al., 2023b)	CLIP ViT-L/14 (Radford et al., 2021)	Vicuna-7B	30.79
DeepSeek-VL-Chat (Lu et al., 2024)	(SigLIP-L (Zhai et al., 2023) and ViTDet (Li et al., 2022))	DeepSeek-LLM-7B	31.99
Yi-VL (Young et al., 2024)	Openclip ViT-H/14 (Ilharco et al., 2021)	Yi-6B	29.96
Gemini-Pro-Vision	-	-	30.46
GPT-4V (xxxx, 2023)	-	-	37.98

Bold values indicate the best performance

incorrect plans, such as concluding that the next action is to “Add yogurt to meat,” when this step has already been completed. This case study highlights a key area for improvement in MLLMs’ processing of visual cues in the context of goal-oriented task execution.

Key Findings from Common Failure Cases We further provide some common failure cases of the evaluated MLLMs in Figure 8. By employing qualitative and quantitative analysis, we summarize three key findings for explaining the limitations of current MLLMs in human-level planning:

(1) Insufficient integration of visual modality: Most MLLMs, adapted from Large Language Models, excel in semantic reasoning but struggle with real-world observations necessary for human-level planning. Consequently, they often select options semantically closer to the task goal while ignoring visual cues (see Failure Type-1 in Figure 8). When we categorized EgoPlan-Val instances into three difficulty levels based on semantic similarity (calculated by BertScore Zhang et al. 2019) between the task goal and the ground-truth answer, the top three models perform significantly better on easier questions with higher goal-answer similarity (Figure 9).

(2) Omitting key state changes in task progress: Long-horizon tasks require models to monitor state changes of key objects, necessitating both long-context memory and high-resolution visual observation processing. As task progress lengthens (measured by the number of historical actions), it becomes harder to capture state changes, and the model performance generally decreases (Figure 10). For example,

in Failure Type-2 (Figure 8), the model plans to “pour olive oil on salad” without recognizing that this action has already been completed.

(3) Inadequate application of world knowledge: Human-level planning also relies on world knowledge. In Failure Type-3 (Figure 8), to “store leftover food,” “closing the container” is essential for preserving freshness before “putting it somewhere”. However, current MLLMs struggle to integrate their inherent world knowledge into the planning process as seamlessly as humans.

5 Enhancing Human-Level Planning through Instruction Tuning

Given the suboptimal performance of the evaluated MLLMs on EgoPlan-Val, we investigate enhancing the human-level planning capabilities of MLLMs through instruction-tuning. Specifically, we construct an instruction-tuning dataset, EgoPlan-IT, to enable MLLMs to effectively learn human-level planning experience from real-world videos. In this section, we first describe the data preparation process in Sec. 5.1. Then we introduce the framework of a baseline MLLM in 5.2 and the training objectives in Sec. 5.3. In Sec. 5.4, we demonstrate the effectiveness of our data and instruction-tuning strategy through experiments.

5.1 Data Preparation

Construction of EgoPlan-IT. Following the pipeline in Sec. 3, we additionally tailor an instruction-tuning dataset

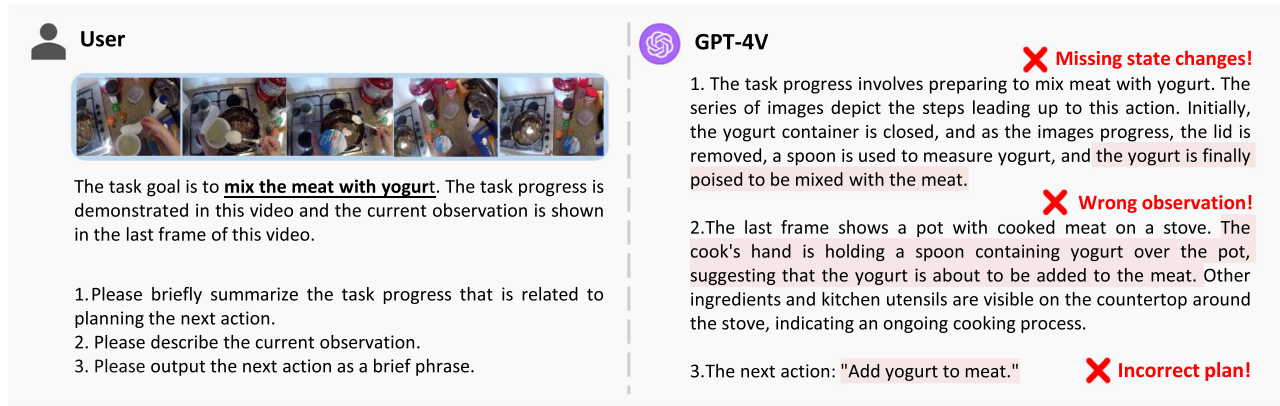


Fig. 7 A case study with GPT-4V, where the model fails to detect the state changes of key objects in task progress, and struggles to recognize the target object in current observation, leading to the wrong prediction of action plan

EgoPlan-IT, which contains 50K QA pairs for human-level planning, as shown in the bottom half of Figure 2. To assess the robustness of our instruction-tuning approach, we exclusively utilize videos from Epic-Kitchens (Damen et al., 2022) for the creation of this dataset. This allows us to assess the enhanced model's performance on both the in-domain (Epic-Kitchens) subset, comprising 2,432 samples from EgoPlan-Val, and the out-of-domain (Ego4D) subset, consisting of 923 samples from EgoPlan-Val. In order to increase the efficiency for constructing this dataset, we do not include human annotations in this process.

Auxiliary Data for Action Recognition. To alleviate the issue of missing key state changes during task planning, we incorporate auxiliary action recognition data during instruction tuning, which helps the model better track task progress. Specifically, this auxiliary dataset is constructed based on the visual inputs of the 50K training samples from EgoPlan-IT. For example in the bottom half of Figure 2, the instruction of the auxiliary sample would be a simple query “Can you enumerate the actions in the video, describing each with a short verb-noun combination?”, and the response is the concatenation of corresponding action narrations such as “Open yogurt, scoop yogurt, put yogurt on meat, put down yogurt.”

Increasing Instruction Diversity. In order to preserve the broad world knowledge inside the pre-trained base model, we co-train the model with diverse general-domain multimodal instruction data. Specifically, we include an additional collection of 164K instruction data, which integrates the 150K image-instruction dataset from LLaVA (Liu et al., 2023a), the 3K image-detail-description dataset from MiniGPT-4 (Zhu et al., 2023) and the 11K video-instruction dataset from VideoChat (Li et al., 2023).

5.2 Model Framework

We use Video-LLaMA (Zhang et al., 2023) as a baseline MLLM for investigation. As illustrated in Figure 11, the

vision processor of Video-LLaMA comprises four components, including an Image Encoder to extract features from video frames, a position embedding layer for adding temporal information, a Video Q-former for aggregating frame representations, and a linear layer for producing video embedding vectors that match the dimensions of LLM token embeddings.

To enhance the model's ability to process visual inputs for human-level planning, we organize historical observations based on actions and retain the most recent $N(=4)$ clips. For each action segment, we choose $K(=8)$ uniformly spaced keyframes, creating an aggregated video representation. To emphasize the current observation, we treat it as a separate single-frame clip, extracting a unique representation for it. The token embeddings for placeholders $\langle \text{Clip}_i \rangle$ and $\langle \text{Observation} \rangle$ in the textual prompt are replaced with the embeddings corresponding to the i -th action segment and the current observation, respectively, within the LLM.

5.3 Training Objectives

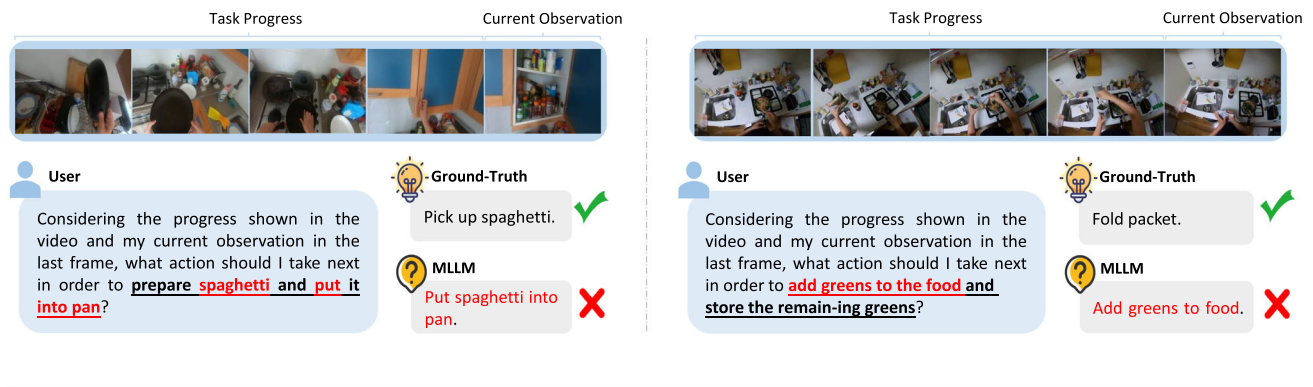
To fine-tune the model, we mix the 50K specialized EgoPlan-IT data together with the 50K auxiliary data for Action Recognition and the 164K diverse instruction data. The autoregressive loss function is defined as:

$$L_{ft} = - \sum_t \log P(a_t | \mathbf{v}, \mathbf{l}, \mathbf{a}_{<t}; \Theta), \quad (1)$$

where $\mathbf{v}$ and $\mathbf{l}$ denote the given visual observations and the language instruction, and $\mathbf{a}_{<t}$ signifies the sequence of answer tokens preceding the current prediction token a_t . Θ is the trainable model parameters with LoRA (Hu et al., 2021).

In order to enhance the model in integrating visual modality during task planning, we add a contrastive loss that encourages the model to distinguish correct action plans from incorrect ones within varying visual contexts, inspired by

Failure Type-1: Insufficient integration of visual modality.



Type-2: Omitting key state changes in task progress.



Type-3: Inadequate application of world knowledge.

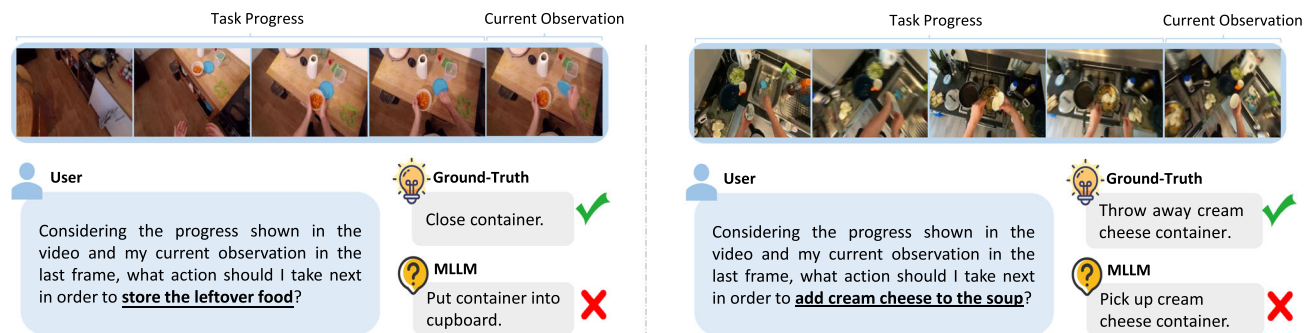


Fig. 8 Common failure cases of the evaluated MLLMs

RRHF (Yuan et al., 2023). The overall fine-tuning loss function is shown as follows:

$$L = L_{ft} + L_{ct}, \quad (2)$$

$$L_{ct} = \max(0, s_{neg} - s_{pos}), \quad (3)$$

$$s = \frac{\sum_t \log P(a_t | v, l, a_{<t}; \Theta)}{||a||}. \quad (4)$$

Here, for each instruction-following pair, a random action plan distinct from the ground-truth answer is selected under the same task goal to serve as a 'negative' answer. The model is constrained to ensure that its score for the 'positive' answer (s_{pos}) exceeds that for the 'negative' answer (s_{neg}). This operation encourages a more balanced consideration of both the language and visual modalities.

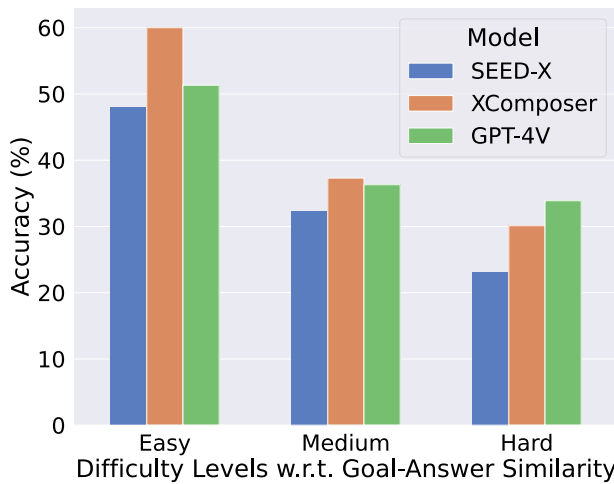


Fig. 9 The influence of goal-answer similarity on the top-3 model performance

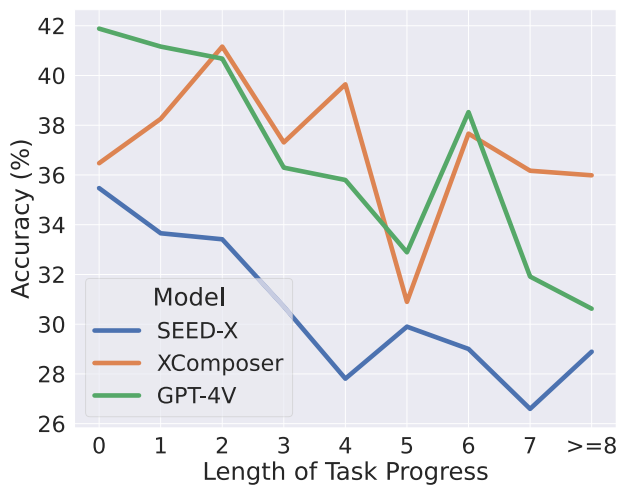


Fig. 10 The influence of task progress length on the top-3 model performance

5.4 Experiments

Implementation Details. The baseline MLLM is built upon the Video-LLaMA (LLaMA2-Chat-7B) checkpoint which has undergone the second training stage². We further fine-tune the model on our data for 2 epochs using a batch size of 16 on 8 V100 GPUs, which takes 0.5 days. During this process, the Image Encoder is frozen and the language model is tuned using LoRA strategy, while the remaining components of the model are fully trainable.

Main Results. The results in Table 5 clearly indicate that the baseline model (Video-LLaMA) tuned on our data outperforms the previous state-of-the-art model, GPT-4V, by a

Table 5 Performance of the baseline model (Video-LLaMA) tuned on EgoPlan-IT. In-Dom. and Out-Dom. are short for in-domain and out-of-domain respectively

Model	In-Dom.	Out-Dom.	All
GPT-4V	38.40	36.90	37.98
Video-LLaMA	27.88	30.44	28.58
Enhanced Video-LLaMA	54.65	44.42	51.83
w/o LoRA	53.58	42.58	50.55
w/o Contrastive Loss	52.14	40.52	48.94
w/o Diverse Instructions	53.95	43.12	50.97
w/o Action Recognition	52.34	44.31	50.13

Bold values indicate the best performance

considerable margin. Notably, it shows a remarkable 23.25% increase in accuracy over its vanilla version across the overall validation set. Furthermore, the enhanced model demonstrates robust domain transfer capabilities, outperforming the vanilla version by 13.98% on the out-of-domain validation subset. This indicates that the planning experience gained through instruction-tuning on our EgoPlan-IT data is effectively transferable to new environments.

Ablations. Our analysis of the ablated versions of the enhanced model reveals several key insights. Notably, employing the LoRA strategy, integrating contrastive loss, diversifying instructions, and fine-tuning with auxiliary action recognition data consistently enhance performance across domains. Contrastive loss contributes the most to the model's improvement. Furthermore, fine-tuning the model with in-domain action recognition auxiliary data boosts in-domain performance but does not significantly improve domain transferability.

6 Extending to Multi-Step Task Planning

While our benchmark primarily focuses on single-step action prediction, we acknowledge the importance of evaluating multi-step planning capabilities for achieving human-level task planning. This section discusses our design considerations for single-step evaluation and introduces an extension to multi-step planning.

6.1 Grounding of Single-Step Predictions

Our evaluation questions have undergone rigorous human verification to ensure that the correct next action is uniquely defined and tightly grounded in the current task goal, task progress, and real-time observation. As illustrated in Figure 12, the future action sequence may include multiple steps, but the most appropriate immediate next action is clearly determined by the current context.

² <https://huggingface.co/DAMO-NLP-SG/Video-LLaMA-2-7B-Fine-tuned>

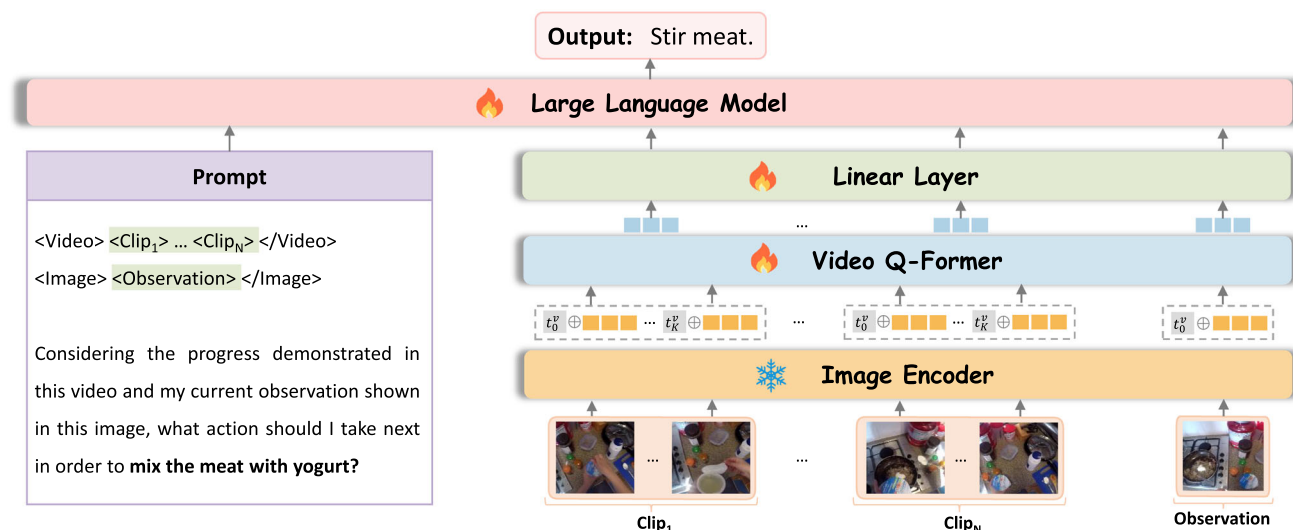


Fig. 11 The diagram of enhancing the baseline model Video-LLaMA with EgoPlan-IT

Task Progress
Current Observation

Question (1-step): Considering the progress shown in the video and my current observation in the last frame, what action should I take next in order to make coffee with milk (**Task Goal**)?

Candidate Options:

A. pick milk bottle | B. open fridge | C. put down mug | D. ☒ close milk bottle

Answer: D

Question (multi-step): Considering the progress shown in the video and my current observation in the last frame, what is the correct sequence of actions I should follow next in order to make coffee with milk (**Task Goal**)?

Candidate Options:

A. ☒ close milk bottle, open fridge, put milk bottle in fridge door, close fridge

B. open fridge, ~~close fridge~~, ~~open milk bottle~~, ~~pour milk from bottle into mug~~

C. ~~close fridge~~, close milk bottle, open fridge, put milk bottle in fridge door

D. close milk bottle, ~~close fridge~~, ~~open milk bottle~~, put milk bottle in fridge door

Answer: A

Fig. 12 Comparison of single- vs. multi-step planning questions with identical visual inputs

For instance, consider a scenario where a person is standing at the countertop with an open milk bottle in hand, and the fridge is located elsewhere in the kitchen. The future action sequence may include steps such as “close milk bottle” to prevent spoilage, followed by “open fridge, put milk in fridge door, close fridge” to store the milk properly. However, the most appropriate immediate next action in the current context is clearly “close milk bottle.” This action is the most reasonable and feasible next step given the immediate environment state and task progress. This example highlights how

our one-step questions are carefully designed to reflect the immediate, contextually grounded next action rather than an abstract or ambiguous plan.

Moreover, this single-step framework intentionally mitigates the error accumulation problem inherent in multi-step prediction. By requiring the ground-truth next action to be closely tied to the current environment state, our benchmark simulates realistic decision-making scenarios faced by intelligent humanoid assistants. For example, a robot tasked with making coffee should close the milk bottle immediately after

Table 6 MLLM Performance on EgoPlan-Val-Multi, evaluated by explicit candidate option selection

Model	Visual Encoder	LLM	Acc%
Random	-	-	25.84
BLIP-2 (Li et al., 2023)	EVA-CLIP ViT-g/14 (Fang et al., 2023)	Flan-T5-XL	36.14
InstructBLIP (Dai et al., 2023)	EVA-CLIP ViT-g/14 (Fang et al., 2023)	Flan-T5-XL	30.90
InstructBLIP Vicuna (Dai et al., 2023)	EVA-CLIP ViT-g/14 (Fang et al., 2023)	Vicuna-7B	24.72
Valley (Luo et al., 2023)	CLIP ViT-L/14 (Radford et al., 2021)	LLaMA-13B	24.91
Qwen-VL-Chat (Bai et al., 2023)	Openclip ViT-bigG (Ilharco et al., 2021)	Qwen-7B	30.33
LLaVA-1.5 (Liu et al., 2023b)	CLIP ViT-L/14 (Radford et al., 2021)	Vicuna-7B	32.56
DeepSeek-VL-Chat (Lu et al., 2024)	(SigLIP-L (Zhai et al., 2023) and ViTDet (Li et al., 2022))	DeepSeek-LLM-7B	27.34
Yi-VL (Young et al., 2024)	Openclip ViT-H/14 (Ilharco et al., 2021)	Yi-6B	33.52
GPT-4o (Hurst et al., 2024)	-	-	43.45

Bold values indicate the best performance

pouring, rather than prematurely navigating to the fridge. If the action fails (e.g., drops the bottle cap), the subsequent correct action would be to pick up the cap and retry closing the bottle, reflecting a dynamic and adaptive planning process. This stepwise evaluation thus aligns well with practical applications where each action depends on the current state and previous outcomes, reducing compounding errors and better reflecting real-time task planning.

6.2 From Single Steps to Sequential Planning: EgoPlan-Val-Multi

Building upon the solid foundation of single-step evaluation, we extend our benchmark with a new set of 534 multi-step planning questions, denoted as EgoPlan-Val-Multi, to enable a more comprehensive evaluation. These questions maintain our rigorous evaluation standards while assessing models' ability to predict sequences of actions.

As shown in Figure 12, the multi-step questions are formulated as multiple-choice problems to maintain evaluation consistency, avoiding the complexities of free-form sequence generation and matching. The negative choices include: 1) shuffled versions of the ground-truth action sequence, 2) sequences reflecting only historical task progress, and 3) randomly sampled actions from the entire task action space.

Table 6 presents evaluation results of 10 representative MLLMs on this extended benchmark. The results demonstrate that current models struggle with multi-step planning, with even the best-performing model achieving only 43.45% accuracy, highlighting substantial room for improvement in this challenging aspect of task planning.

6.3 Limitations and Future Directions

While our multi-step evaluation provides a meaningful advancement, it does not yet fully capture the complexities of real-world task planning. Evaluating multi-step plans is inherently more challenging due to 1) the expanded solution space of compositional action sequences, 2) the greater uncertainty in long-horizon predictions, and 3) the lack of dynamic environment state updates after each action.

A more comprehensive evaluation would involve realistic simulation environments where a humanoid assistant executes predicted action sequences in real time, allowing assessment of overall task success rates. This represents an important direction for future work in human-level task planning evaluation.

7 Conclusion

In this work, we introduce EgoPlan-Bench, a sophisticated benchmark to evaluate Multimodal Large Language Models (MLLMs) in human-level planning, where a model is expected to plan step-by-step actions by considering task progress, current visual observations, and open-form task goals. The evaluation results of various models reveal that current MLLMs have not yet evolved into generalized human-level planners. We further construct EgoPlan-IT, an instruction-tuning dataset, to facilitate the learning of task planning experience from human videos. The model tuned on EgoPlan-IT exhibits a significant performance improvement on our benchmark. We have released all the codes

and datasets, and continuously maintain a benchmark leaderboard to promote future research.

Acknowledgements The research work described in this paper was conducted in the JC STEM Lab of Autonomous Intelligent Systems funded by The Hong Kong Jockey Club Charities Trust.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. and others (2020). Language models are few-shot learners. *Advances in neural information processing systems* 33, 1877–1901.
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., and others (2022). Training language models to follow instructions with human feedback. *NeurIPS*.
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., and others (2023). Llama: Open and efficient foundation language models. [arXiv:2302.13971](https://arxiv.org/abs/2302.13971).
- Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., and others (2023). Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. See <https://vicuna.lmsys.org> (accessed 14 April 2023).
- Arcas, B.A., & Norvig, P. (2023). Artificial general intelligence is already here. *Noema*, October.
- Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y.T., Li, Y., Lundberg, S., and others (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. [arXiv:2303.12712](https://arxiv.org/abs/2303.12712).
- Morris, M.R., Sohl-dickstein, J., Fiedel, N., Warkentin, T., Dafoe, A., Faust, A., Farabet, C., & Legg, S. (2023). Levels of agi: Operationalizing progress on the path to agi. [arXiv:2311.02462](https://arxiv.org/abs/2311.02462).
- Fan, C. (2019). EgoVQA-an egocentric video question answering benchmark dataset. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 0–0.
- Jia, B., Lei, T., Zhu, S.-C., & Huang, S. (2022). EgoTaskQA: Understanding human tasks in egocentric videos. *Advances in Neural Information Processing Systems*, 35, 3343–3360.
- Li, B., Ge, Y., Ge, Y., Wang, G., Wang, R., Zhang, R., & Shan, Y. (2023). Seed-bench-2: Benchmarking multimodal large language models. [arXiv:2311.17092](https://arxiv.org/abs/2311.17092).
- Damen, D., Doughty, H., Farinella, G. M., Furnari, A., Ma, J., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., & Wray, M. (2022). Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision (IJCV)*, 130, 33–55.
- Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., and others (2022). Ego4d: Around the world in 3,000 hours of egocentric video. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18995–19012.
- Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, E., Wang, X., Dehghani, M., Brahma, S., and others (2022). Scaling instruction-finetuned language models. [arXiv:2210.11416](https://arxiv.org/abs/2210.11416).
- Zhang, A., Fei, H., Yao, Y., Ji, W., Li, L., Liu, Z., & Chua, T.-S. (2023). Transfer visual prompt generator across llms. [arXiv:abs/2304.50127](https://arxiv.org/abs/2304.50127).
- Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., Wang, J., Hu, A., Shi, P., Shi, Y., and others (2023). mplug-owl: Modularization empowers large language models with multimodality. [arXiv:2304.14178](https://arxiv.org/abs/2304.14178).
- Li, B., Zhang, Y., Chen, L., Wang, J., Yang, J., & Liu, Z. (2023). Otter: A multi-modal model with in-context instruction tuning. [arXiv:2305.03726](https://arxiv.org/abs/2305.03726).
- Gong, T., Lyu, C., Zhang, S., Wang, Y., Zheng, M., Zhao, Q., Liu, K., Zhang, W., Luo, P., & Chen, K. (2023). MultiModal-GPT: A Vision and Language Model for Dialogue with Humans.
- Awadalla, A., Gao, I., Gardner, J., Hessel, J., Hanafy, Y., Zhu, W., Marathe, K., Bitton, Y., Gadre, S., Sagawa, S., Jitsev, J., Kornblith, S., Koh, P.W., Ilharco, G., Wortsman, M., & Schmidt, L. (2023). Openflamingo: An open-source framework for training large autoregressive vision-language models. [arXiv:2308.01390](https://arxiv.org/abs/2308.01390).
- Gao, P., Han, J., Zhang, R., Lin, Z., Geng, S., Zhou, A., Zhang, W., Lu, P., He, C., Yue, X., Li, H., & Qiao, Y. (2023). Llama-adapter v2: Parameter-efficient visual instruction model. [arXiv:2304.15010](https://arxiv.org/abs/2304.15010).
- Peng, Z., Wang, W., Dong, L., Hao, Y., Huang, S., Ma, S., & Wei, F. (2023). Kosmos-2: Grounding multimodal large language models to the world. [arXiv:2306.14824](https://arxiv.org/abs/2306.14824).
- Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., & Zhou, J. (2023). Qwen-vl: A frontier large vision-language model with versatile abilities. [arXiv:2308.12966](https://arxiv.org/abs/2308.12966).
- Zhang, P., Wang, X.D.B., Cao, Y., Xu, C., Ouyang, L., Zhao, Z., Ding, S., Zhang, S., Duan, H., Yan, H. and others (2023). Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. [arXiv:2309.15112](https://arxiv.org/abs/2309.15112).
- Wang, G., Ge, Y., Ding, X., Kankanhalli, M., & Shan, Y. (2023). What makes for good visual tokenizers for large language models? [arXiv:2305.12223](https://arxiv.org/abs/2305.12223).
- Li, J., Li, D., Savarese, S., & Hoi, S. (2023). Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *ICML*.
- Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., & Hoi, S. (2023). Instructblip: Towards general-purpose vision-language models with instruction tuning. [arXiv:2305.06500](https://arxiv.org/abs/2305.06500).
- Zhu, D., Chen, J., Shen, X., Li, X., & Elhoseiny, M. (2023). Minigpt-4: Enhancing vision-language understanding with advanced large language models. [arXiv:2304.10592](https://arxiv.org/abs/2304.10592).
- Liu, H., Li, C., Wu, Q., & Lee, Y.J. (2023). Visual instruction tuning. [arXiv:2304.08485](https://arxiv.org/abs/2304.08485).
- Liu, H., Li, C., Li, Y., & Lee, Y.J. (2023). Improved baselines with visual instruction tuning. [arXiv:2310.03744](https://arxiv.org/abs/2310.03744).
- Ye, Q., Xu, H., Ye, J., Yan, M., Hu, A., Liu, H., Qian, Q., Zhang, J., Huang, F., & Zhou, J. (2023). mPLUG-Owl2: Revolutionizing Multi-modal Large Language Model with Modality Collaboration.
- Lu, H., Liu, W., Zhang, B., Wang, B., Dong, K., Liu, B., Sun, J., Ren, T., Li, Z., Sun, Y. and others (2024). Deepseek-vl: towards real-world vision-language understanding. [arXiv:2403.05525](https://arxiv.org/abs/2403.05525).
- Young, A., Chen, B., Li, C., Huang, C., Zhang, G., Zhang, G., Li, H., Zhu, J., Chen, J., Chang, J. and others (2024). Yi: Open foundation models by 01. ai. [arXiv:2403.04652](https://arxiv.org/abs/2403.04652).
- Zhang, H., Li, X., & Bing, L. (2023). Video-llama: An instruction-tuned audio-visual language model for video understanding. [arXiv:2306.02858](https://arxiv.org/abs/2306.02858).

- Li, K., He, Y., Wang, Y., Li, Y., Wang, W., Luo, P., Wang, Y., Wang, L., & Qiao, Y. (2023). Videochat: Chat-centric video understanding. [arXiv:2305.06355](#).
- Maaz, M., Rasheed, H., Khan, S., & Khan, F.S. (2023). Video-chatgpt: Towards detailed video understanding via large vision and language models. [arXiv:2306.05424](#).
- Ge, Y., Ge, Y., Zeng, Z., Wang, X., & Shan, Y. (2023). Planting a seed of vision in large language model. [arXiv:2307.08041](#).
- Ge, Y., Zhao, S., Zeng, Z., Ge, Y., Li, C., Wang, X., & Shan, Y. (2023). Making llama see and draw with seed tokenizer. [arXiv:2310.01218](#).
- Ge, Y., Zhao, S., Zhu, J., Ge, Y., Yi, K., Song, L., Li, C., Ding, X., & Shan, Y. (2024). Seed-x: Multimodal models with unified multi-granularity comprehension and generation. [arXiv:2404.14396](#).
- Gpt-4v(ision) system card. (2023). <https://api.semanticscholar.org/CorpusID:263218031>
- Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A. and others (2023). Gemini: a family of highly capable multimodal models. [arXiv:2312.11805](#).
- Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Yang, J., Zheng, X., Li, K., Sun, X., and others (2023). Mme: A comprehensive evaluation benchmark for multimodal large language models. [arXiv:2306.13394](#).
- Yin, Z., Wang, J., Cao, J., Shi, Z., Liu, D., Li, M., Huang, X., Wang, Z., Sheng, L., Bai, L., et al. (2024). Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. *Advances in Neural Information Processing Systems*, 36, 26650–26685.
- Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z. and others (2023). Mmbench: Is your multi-modal model an all-around player? [arXiv:2307.06281](#).
- Yu, W., Yang, Z., Li, L., Wang, J., Lin, K., Liu, Z., Wang, X., & Wang, L. (2023). Mm-vet: Evaluating large multimodal models for integrated capabilities. [arXiv:2308.02490](#).
- Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y. and others (2023). Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. [arXiv:2311.16502](#).
- Xu, P., Shao, W., Zhang, K., Gao, P., Liu, S., Lei, M., Meng, F., Huang, S., Qiao, Y., & Luo, P. (2023). Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. [arXiv:2306.09265](#).
- Ying, K., Meng, F., Wang, J., Li, Z., Lin, H., Yang, Y., Zhang, H., Zhang, W., Lin, Y., Liu, S. and others (2024). Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi. [arXiv:2404.16006](#).
- Cheng, S., Guo, Z., Wu, J., Fang, K., Li, P., Liu, H., & Liu, Y. (2023). Can vision-language models think from a first-person perspective? [arXiv:2311.15596](#).
- Chen, L., Zhang, Y., Ren, S., Zhao, H., Cai, Z., Wang, Y., Wang, P., Liu, T., & Chang, B. (2023). Towards end-to-end embodied decision making via multi-modal large language model: Explorations with gpt4-vision and beyond. [arXiv:2310.02071](#).
- Sener, F., & Yao, A. (2019). Zero-shot anticipation for instructional activities. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 862–871.
- AbuFarha, Y., Ke, Q., Schiele, B., & Gall, J. (2021). Long-term anticipation of activities with cycle consistency. In: Pattern Recognition: 42nd DAGM German Conference, DAGM GCPR 2020, Tübingen, Germany, September 28–October 1, 2020, Proceedings 42, pp. 159–173. Springer.
- Furnari, A., Battiato, S., Grauman, K., & Farinella, G. M. (2017). Next-active-object prediction from egocentric videos. *Journal of Visual Communication and Image Representation*, 49, 401–411.
- Liu, S., Tripathi, S., Majumdar, S., & Wang, X. (2022). Joint hand motion and interaction hotspots prediction from egocentric videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3282–3292.
- Lotter, W., Kreiman, G., & Cox, D. (2016). Deep predictive coding networks for video prediction and unsupervised learning. In: International Conference on Learning Representations.
- Villegas, R., Yang, J., Hong, S., Lin, X., & Lee, H. (2017). Decomposing motion and content for natural video sequence prediction. [arXiv:1706.08033](#).
- Mendonca, R., Bahl, S., & Pathak, D. (2023). Structured world models from human videos. [arXiv:2308.10901](#).
- Lin, X., Petroni, F., Bertasius, G., Rohrbach, M., Chang, S.-F., & Torresani, L. (2022). Learning to recognize procedural activities with distant supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13853–13863.
- Zhong, Y., Yu, L., Bai, Y., Li, S., Yan, X., & Li, Y. (2023). Learning procedure-aware video representation from instructional videos and their narrations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14825–14835.
- Chang, C.-Y., Huang, D.-A., Xu, D., Adeli, E., Fei-Fei, L., & Niebles, J.C. (2020). Procedure planning in instructional videos. In: European Conference on Computer Vision, pp. 334–350. Springer.
- Bi, J., Luo, J., & Xu, C. (2021). Procedure planning in instructional videos via contextual modeling and model-based policy learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 15611–15620.
- Sun, J., Huang, D.-A., Lu, B., Liu, Y.-H., Zhou, B., & Garg, A. (2022). Plate: Visually-grounded planning with transformers in procedural tasks. *IEEE Robotics and Automation Letters*, 7(2), 4924–4930.
- Zhao, H., Hadji, I., Dvornik, N., Derpanis, K.G., Wildes, R.P., & Jepson, A.D. (2022). P3iv: Probabilistic procedure planning from instructional videos with weak supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2938–2948.
- Patel, D., Eghbalzadeh, H., Kamra, N., Iuzzolino, M.L., Jain, U., & Desai, R. (2023). Pretrained language models as visual planners for human assistance. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 15302–15314.
- Zhukov, D., Alayrac, J.-B., Cinbis, R.G., Fouhey, D., Laptev, I., & Sivic, J. (2019). Cross-task weakly supervised learning from instructional videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3537–3545.
- Tang, Y., Ding, D., Rao, Y., Zheng, Y., Zhang, D., Zhao, L., Lu, J., & Zhou, J. (2019). Coin: A large-scale dataset for comprehensive instructional video analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1207–1216.
- Lai, B., Dai, X., Chen, L., Pang, G., Reh, J.M., & Liu, M. (2023). Lego: Learning egocentric action frame generation via visual instruction tuning. [arXiv:2312.03849](#).
- Pirsiavash, H., & Ramanan, D. (2012). Detecting activities of daily living in first-person camera views. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2847–2854. IEEE.
- Li, Y., Ye, Z., & Reh, J.M. (2015). Delving into egocentric actions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 287–295.
- Sigurdsson, G.A., Gupta, A., Schmid, C., Farhadi, A., & Alahari, K. (2018). Charades-ego: A large-scale dataset of paired third and first person videos. [arXiv:1804.09626](#).
- Song, Y., Byrne, E., Nagarajan, T., Wang, H., Martin, M., & Torresani, L. (2024). Ego4d goal-step: Toward hierarchical understanding of procedural activities. *Advances in Neural Information Processing Systems*, 36, 38863–38886.
- OpenAI: GPT-4 Technical Report (2023)
- Lin, K.Q., Wang, J., Soldan, M., Wray, M., Yan, R., XU, E.Z., Gao, D., Tu, R.-C., Zhao, W., Kong, W. and others (2022). Egocentric video-

- language pretraining. *Advances in Neural Information Processing Systems* **35**, 7575–7586.
- Brohan, A., Chebotar, Y., Finn, C., Hausman, K., Herzog, A., Ho, D., Ibarz, J., Irpan, A., Jang, E., Julian, R. and others (2023). Do as i can, not as i say: Grounding language in robotic affordances. In: *Conference on Robot Learning*, pp. 287–318. PMLR
- Lin, S., Hilton, J., & Evans, O. (2021). Truthfulqa: Measuring how models mimic human falsehoods. [arXiv:2109.07958](https://arxiv.org/abs/2109.07958).
- Luo, R., Zhao, Z., Yang, M., Dong, J., Qiu, M., Lu, P., Wang, T., & Wei, Z. (2023). Valley: Video assistant with large language model enhanced ability. [arXiv:2306.07207](https://arxiv.org/abs/2306.07207).
- Wang, W., Lv, Q., Yu, W., Hong, W., Qi, J., Wang, Y., Ji, J., Yang, Z., Zhao, L., Song, X. and others (2023). Cogvlm: Visual expert for pretrained language models. [arXiv:2311.03079](https://arxiv.org/abs/2311.03079).
- Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Liu, Y., Wang, Z., Xu, J., Chen, G., Luo, P. and others (2024). Mvbench: A comprehensive multi-modal video understanding benchmark. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22195–22206.
- Fang, Y., Wang, W., Xie, B., Sun, Q., Wu, L., Wang, X., Huang, T., Wang, X., & Cao, Y. (2023). Eva: Exploring the limits of masked visual representation learning at scale. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19358–19369.
- Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J. and others (2021). Learning transferable visual models from natural language supervision. In: *International Conference on Machine Learning*, pp. 8748–8763. PmLR.
- Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., & Schmidt, L. OpenCLIP. If you use this software, please cite it as below. <https://doi.org/10.5281/zenodo.5143773>
- Zhai, X., Mustafa, B., Kolesnikov, A., & Beyer, L. (2023). Sigmoid loss for language image pre-training. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11975–11986.
- Li, Y., Mao, H., Girshick, R., & He, K. (2022). Exploring plain vision transformer backbones for object detection. In: *European Conference on Computer Vision*, pp. 280–296. Springer.
- Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., & Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. In: *International Conference on Learning Representations*.
- Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). Lora: Low-rank adaptation of large language models. [arXiv:2106.09685](https://arxiv.org/abs/2106.09685).
- Yuan, H., Yuan, Z., Tan, C., Wang, W., Huang, S., & Huang, F. (2023). Rrhf: Rank responses to align language models with human feedback. *Thirty-seventh Conference on Neural Information Processing Systems*, 36, 10935–10950.
- Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A. and others (2024). Gpt-4o system card. [arXiv:2410.21276](https://arxiv.org/abs/2410.21276).

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.