# User-in-the-loop Evaluation of Multimodal LLMs for Activity Assistance

Mrinal Verghese*
Carnegie Mellon University
mverghes@andrew.cmu.edu

Brian Chen*
Samsung Research America
bc2754@columbia.edu

Hamid Eghbalzadeh
Meta Reality Labs Research
heghbalz@meta.com

Tushar Nagarajan
Meta Fundamental AI Research
tusharn@meta.com

Ruta Desai
Meta Fundamental AI Research
rutadesai@meta.com

## Abstract

*Our research investigates the capability of modern multimodal reasoning models, powered by Large Language Models (LLMs), to facilitate vision-powered assistants for multi-step daily activities. Such assistants must be able to 1) encode relevant visual history from the assistant's sensors, e.g., camera, 2) forecast future actions for accomplishing the activity, and 3) replan based on the user in the loop. To evaluate the first two capabilities, grounding visual history and forecasting in short and long horizons, we conduct benchmarking of two prominent **classes** of multimodal LLM approaches – Socratic Models [46] and Vision Conditioned Language Models (VCLMs) [31] on video-based action anticipation tasks using offline datasets. These offline benchmarks, however, do not allow us to close the loop with the user, which is essential to evaluate the replanning capabilities and measure successful activity completion in assistive scenarios. To that end, we conduct a first-of-its-kind user study, with 18 participants performing 3 different multi-step cooking activities while wearing an egocentric observation device called Aria [37] and following assistance from multimodal LLMs. We find that the Socratic approach outperforms VCLMs in both offline and online settings. We further highlight how grounding long visual history, common in activity assistance, remains challenging in current models, especially for VCLMs, and demonstrate that offline metrics do not indicate online performance.*

## 1. Introduction

Imagine a vision-powered assistant capable of empowering its users in multi-step daily activities like cooking, assembling, etc. by detecting mistakes and recommending corrections. Two fundamental capabilities of such an assistant are a) the ability to understand task-relevant steps and progress accomplished by the user from the *past* visual observations e.g., video [33] and b) the ability to recommend the next actions the user should take by forecasting and planning of *future* actions [10, 15]. In addition to encoding history and forecasting, such assistants must also account for the user in the loop and re-plan on the fly to ensure successful task execution. With the advent of various

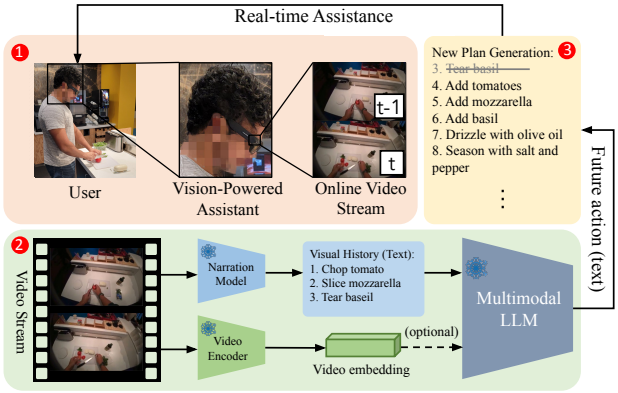



Figure 1. **Online deployment and evaluation of multimodal LLMs for vision-powered assistive systems.** 1) Users equipped with Aria [37], which captures an egocentric video stream of their actions, perform multi-step activities while following assistance from multimodal LLMs. 2) These models encode the video stream using text-based representations and optionally vision embeddings to predict future actions to complete the activity. 3) These actions are replanned during execution to provide real-time assistance.

vision-language models powered by modern-day LLMs, a natural question is how these models fare towards the aforementioned capabilities. To that end, our work makes two contributions. First, we perform benchmarking on offline datasets to understand a) the effectiveness of different approaches for grounding visual history in multimodal LLMs and b) the forecasting capabilities of these approaches in short and long horizons. Second, we test these approaches in online settings to understand whether the performance in offline experiments translates to real-world assistive scenarios with a user in the loop. To the best of our knowledge, our work is the *first* to perform such an online evaluation.

While a plethora of multimodal LLMs exist today [7, 8, 16, 22, 24, 26, 29, 31, 47, 48], they can be broadly differentiated into two categories based on their approach to grounding multimodal inputs – *Socratic Models* and *Vision-conditioned Language Models (VCLMs)*. The Socratic approach [46] converts visual history into text using pre-trained VLMs such as action or object detectors or narration models. Figure 1 shows the visual history represented as

text, e.g., "chop tomato", "slice mozzarella", etc., obtained from a video narration model for a Caprese salad-making activity. On the other hand, VCLMs such as Flamingo [1] or LLaVA [27] etc. can embed the visual information as continuous embeddings along with text tokens. Specifically, VCLMs split up the available input tokens of their inbuilt LLMs into continuous embeddings from a pretrained vision encoder and text tokens. While VCLMs vary in their training data, underlying LLMs, and vision encoders (Appendix 10 provides a detailed overview), they all follow a similar general architecture [8, 24]. Such implicit representation of visual information through continuous embeddings may allow VCLMs to capture details that are hard to encode in text. However, current VCLMs only process a limited number of frames from a long video and tend to uniformly sample these frames [24]. Consequently, they may miss keyframes and crucial details pertaining to activity assistance. A natural question then is – *which of these two approaches is more effective for activity assistance?*

To that end, our offline benchmarking evaluates these multimodal LLM approaches using existing video-based action anticipation and visual planning tasks that are representative of challenges in real-world activity assistance [15, 33]. Specifically, our benchmarks span the spectrum of medium to long visual history and forecasting horizon (Tab. 1). To rigorously compare the Socratic and VCLM approaches, we implement representative models of each using the same LLMs, VLMs, and prompting techniques. We find that Socratic approaches outperform VCLMs in tasks requiring the grounding of a long visual history, regardless of forecasting horizon. The coarse-level information spanning the entire visual history as captured by Socratic is more important than fine-grained information implicitly captured by VCLMs from a limited set of frames for effective planning in tasks with long visual histories. Current VCLMs succeed at providing effective grounding for future action prediction primarily for short to medium visual histories. Further, such implicit vision representation in VCLMs is only helpful for planning with small-sized LLMs.

While these benchmarks evaluate SOTA models in offline settings, it is unclear how such models would perform in online settings where the user is actually performing the activity. Unlike offline datasets, online settings have compute and inference time constraints and require accounting for the user in the loop. The interaction with the user enables the unique opportunity to evaluate the correctness of predicted actions and replanning efficacy toward successful activity completion, which is impossible with offline videos. Specifically, conventional edit distance or success rate metrics for evaluating action plans in action anticipation and planning benchmarks consider matching a single plan present in a given offline video. Instead, online evaluation enables considering all possible action plans that lead to activity success. Albeit, the model must replan at each step, as the user performs the task while grounding the observed task progress from the incoming untrimmed video. To test the ability of multimodal LLMs in such settings, we conducted a study with 18 participants performing 3 different multi-step cooking activities while wearing an egocentric observation device called Aria [37] (Fig. 1). During activity execution, each participant is instructed to ask for and follow the next action assistance from Socratic and VCLMs at different points in the activity to accomplish it. We find that the Socratic approach outperforms VCLMs even in such online settings. Effective online assistance requires identifying task-relevant steps from long, untrimmed, and unsegmented videos while ignoring distractors. Akin to offline settings, the text-based representations of visual history in Socratic models appear better suited to capture such information in comparison to the implicit representations in VCLMs. Lastly, a head-on comparison between online and offline metrics in our study also highlights that offline metrics such as mean Intersection over Union (mIoU) are an inflated measure of online performance.

## 2. Related Work

**LLMs for multimodal reasoning.** Inspired by the capabilities of LLMs, Zeng et al. [46] pioneered the Socratic approach to leverage LLMs for multimodal reasoning. Recently, Palm [16] and AntGPT [48] have employed a similar approach using LLMs to anticipate future actions in videos. Specifically, they transform videos into text using narration models like BLIP-2, action recognition models [48], or a combination of the two [16]. The LLM is then used to model the text sequence to predict future action in a sentence completion fashion [34]. However, these approaches have only been evaluated on offline video datasets. In this work, we evaluate such models in an online manner by deploying them in a real-world assistance setting.

**Vision-conditioned Language Model (VCLM).** Instead of a text-based representation of vision modality for LLM-based reasoning in vision tasks, another approach is to have a unified model that combines visual and linguistic information by aligning these modalities. Examples of these models include Flamingo [1], OpenFlamingo [4], Palm-E [12], BLIP-2 [21], InstructBLIP [9], LLaVA [27], IDEFICS [19], MiniGPT-4 [51] and many more [13, 14, 20, 28, 38, 45]. These models are generally fine-tuned using large-scale datasets containing multimodal data [23, 27] and are evaluated on image captioning [44] and visual question answering (VQA) tasks [2, 3, 11].

Building on these, VideoLLM [7], AnyMAL [31], VideoChat-Embed [22], Video-ChatGPT [29], Video-LLaVA [24], LLaVA-NeXT [26] and Video-LLaMA [8, 47] fine-tuned VCLMs for a set of video tasks. A nuanced overview of these models can be found in Tab. 13 in Appendix 10. Keeping the sparsity of available annotated data in assistance scenarios in mind, we specifically focus on few-shot VCLMs without any task-specific finetuning for future action prediction in videos. Note that various other multimodal multitask transformer models such as GATO [35] operate on multimodal tokenized input/output from various modalities such as text, image, video, robot actions, etc., for planning. However, we focus on multimodal models that use LLMs as a backbone.

## 3. Multimodal LLM Approaches

Activity assistance requires grounding information from untrimmed video history for future action prediction. If the visual history is appropriately represented, the future action prediction task can be framed as a sentence completion task using an LLM [40, 41]. We evaluate two predominant approaches to represent visual history for such reasoning with LLMs to predict future actions. Figure 2a shows an overview of our Socratic and VCLM models.

**Socratic model.** The main idea behind the Socratic approach is to use pretrained vision-language models (VLMs) to convert non-textual modalities into text for a downstream LLM. One could extract and represent different task-relevant information from the untrimmed video history as text, such as objects, actions, and open-set narrations describing the events in the video. We find that open-set narrations tend to be a superset of various contextual information that could be extracted from video, including objects and actions. Appendix 9 provides this quantitative comparison for LTA. Instead of using objects, actions, and narrations, we choose only narrations from a video narration model to represent visual history as text in our Socratic models, akin to various existing works [16, 46]. As an example, Socratic models would represent visual information corresponding to visual history in a curry-making activity such as "Add oil in pan", "Add onions in pan" etc.

**Vision-conditioned language models (VCLMs).** These models embed the visual modality as continuous tokens that can be passed as input to an LLM along with text tokens. A linear [27, 31] or a non-linear projection layer [33] is fine-tuned to align these continuous tokens with the embedding space of text tokens for a given LLM. Finally, the LLM backbone is often fine-tuned on a multimodal instruction dataset. [8, 24, 26, 31] Thus, unlike Socratic Models, VCLMs can process both embedded visual information and text. Such implicit representation may allow models to capture fine-grained visual information e.g., "state of the fried onions" while making a curry, which might decide if the user should stir more or add the next ingredient.

VCLMs typically split up the available tokens in their input context to their inbuilt LLM into continuous embeddings and text tokens, with continuous embeddings coming from visual encoders. The encoders in current SOTA VCLMs use limited and uniformly sampled frames from input videos for video tasks. Most VCLMs process between 8 to 16 frames [8, 24, 26] which may be ineffective in our benchmarks which require grounding on average >500 frames corresponding to multiple task-relevant steps. To ensure that VCLMs could be applied to our benchmarks, we use both text tokens and continuous embeddings to encode the visual history in our VCLMs.[1] Appendix 9 shows an ablation comparing VCLMs that only use continuous embeddings for the encoding history with those that use both continuous embeddings and text tokens in LTA.

[1]This is in contrast to VCLMs used for image captioning and VQA tasks, where visual information is only encoded using the continuous embeddings [1].

| Task | Dataset | Metric | Visual history | Forecasting horizon | Prediction space |
|---|---|---|---|---|---|
| Visual Planning for Assistance (VPA) [33] | CrossTask [52] | Mean Accuracy↑, Mean IoU↑, Success Rate ↑ | Medium (3-4 actions) | Medium (3-4 actions) | 118 actions |
| Long-term Action Anticipation (LTA) [15] | Ego4D [15] | Edit Distance ↓ | Long (>= 8 actions) | Long (20 actions) | 115 verbs, 478 nouns 3542 actions |

Table 1. **Offline benchmarks.** We consider benchmarks spanning the spectrum of medium to long visual history and forecasting horizon, which capture the needs of real-world vision-powered activity assistants. Note that we only consider feasible actions in our prediction space instead of all possible combinations of verbs and nouns.

## 4. Offline Benchmarks

Our goal is to make progress toward vision-powered assistants that can reason about their user's context from visual input, such as the user's progress in daily activities, and provide relevant recommendations on future actions. Various action anticipation benchmarks previously proposed by the research community also require such reasoning capabilities. Hence, we choose them to evaluate the two prominent categories of SOTA multimodal LLM approaches.

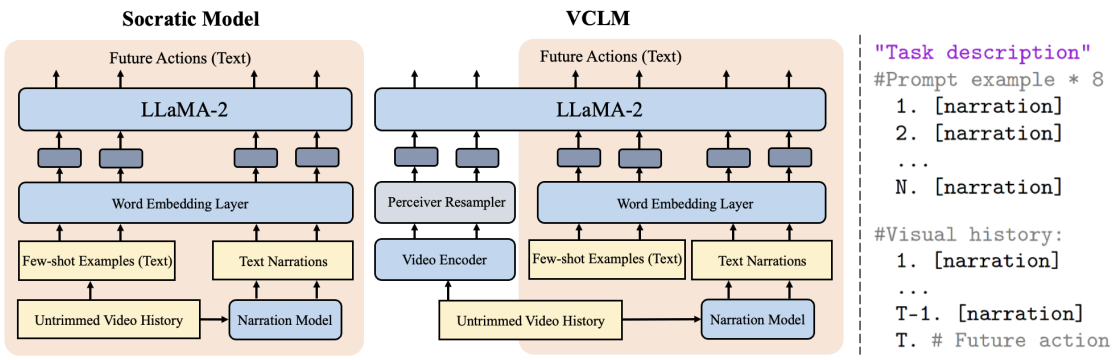
### 4.1. Benchmark Tasks

While a plethora of video-based action anticipation benchmarks exist [32], we choose two representative ones such that they cover the space of medium to long visual history and medium to long forecasting horizon – the settings closest to activity assistance in real-world vision-powered systems[2] Specifically, we choose Long-term action Anticipation (LTA) from Ego4D [15], and Visual Planning for Assistance (VPA) task on the CrossTask dataset from [33]. We blurred faces from the CrossTask videos prior to use. As summarized in Table 1, LTA focuses on predicting a sequence of future actions with a length of $Z = 20$ after grounding a long untrimmed visual history corresponding to approximately 8 or more actions. Compared to LTA, VPA on CrossTask operates on a medium-range untrimmed visual history corresponding to 3-4 actions for medium-horizon forecasting of $Z = 3 \sim 4$ future actions. While not part of the original benchmark, we also look at LTA with $Z = 5$ to help disambiguate the challenges of long-history and long-horizon in prediction. The output is mapped to a closed set of verbs, nouns, and actions, i.e., (verb, noun) pairs in each benchmark. We use the same evaluation metrics as were proposed by the original benchmarks for consistency with prior work. The predicted action sequences in LTA are evaluated using the edit distance. VPA is evaluated with action prediction accuracy at each step (mean accuracy), order-agnostic mean Intersection over Union (mIoU), and a strict order-respecting metric Success Rate for the predicted sequence (defined as in [33]).

### 4.2. Experiment Setup

Figure 2 shows an overview of Socratic and VCLM models used in our experiments. Both our Socratic and VCLM implementations use the same video narration model to encode the visual history as text. For continuous visual repre-

[2]We exclude anticipation benchmarks like EpickKitchens [10] as they consider short visual histories (1s) and are not aligned with real-world activity assistance.




(a) Architecture of Socratic and VCLM (b) Few-shot Examples

Figure 2. **Architecture for Socratic and VCLM with few-shot examples.** We perform few-shot inference of future actions in our benchmark tasks using untrimmed video history as input. Both Socratic and VCLM convert the video history into text narrations using video narration models (as highlighted in the shaded box). Apart from the narrations, VCLM also embeds the video history as continuous embeddings using a video encoder and Perceiver sampler [31]. The output is a sequence of natural language sentences, which are then mapped to a closed set of actions for each benchmark.

sentations, our VCLM model follows the popular projector-based architecture from recent literature [8, 22, 24, 26] and consists of a video encoder [42] attached to a pretrained LLM with an adapter layer. Both Socratic and VCLM use instruction-tuned 13B and 70B Llama2 [41] as their llm backbone, which has a context length of 2048 tokens. While Socratic models use all 2048 of these tokens for text-based representation of history, VCLMs use 256 tokens out of 2048 for continuous visual embeddings of the history. Since our goal is to understand the capability of these SOTA multimodal LLMs for real-world assistive systems where in-domain data is scarce, we consider these models in the in-context regime only. Appendix 10 has more information on Socratic and VCLM model variants and our selected implementations within each of these classes of models.

**Narrations for visual history.** To obtain the text-based representation of visual history for Socratic and VCLMs, we leverage video captioning models. Specifically, we used LaViLa [49]) pre-trained on the Ego4D [15] dataset for Ego4D LTA and video encoder setup following previous work [33] for VPA. We ensure the best performance in terms of grounding for these multimodal LLMs by a) using fine-tuned video narration models and b) by using ground-truth segments of the visual history following previous work [16, 48] in LTA and segments from a finetuned segmentation model following previous work [33] in VPA.

**Prompts for in-context learning.** For each benchmark task, we generate a set of in-context examples by running the narration model on full, segmented videos from the training set. At test time, we use retrieval-based prompting following [16] to select a set of in-context examples that are semantically similar to the narrations from the visual history of the input video. Our final LLM prompt consists of a prompt describing the task of predicting future actions, the set of in-context examples, and the narrations corresponding to the input video's history as shown in Figure 2(b). For VPA, following previous work [33], we also include the goal of the activity in the video in the prompt by concatenating in front of the narration history. All tasks use 8 in-context examples except where the number of tokens exceeds the LLM's context window.

**Mapping free-form predicted sentences to closed-set action classes.** The output of our multimodal LLMs consists of free-form sentences containing predicted future actions. We split predicted sentences based on newline characters and map them onto the closed-set noun and verb classes for each benchmark by finding the closest word in our class label vocabulary to each word in the output sentence based on the cosine similarity in the SentenceBERT embedding space [39].

**Baselines.** We compare VCLMs and Socratic approaches with a set of existing LLM-based and supervised approaches (Tables 2 and 3). We select these baselines based on the availability of code or results in the Ego4D LTA v1 validation set [15] and the VPA task [33]. **LTA baselines.** We include the supervised baseline provided by the Ego4D paper [15] to represent the vision-only approach (not using LM) and an LLM-based baseline VLaMP [33] [3] finetuned on the LTA trainset. Lastly, we consider few-shot LLM-based approaches – AntGPT [48] and Palm [16], which have shown strong performance on the Ego4D LTA task. AntGPT utilizes the activity goal inferred by Llama2-Chat-13B [41] and the history of recognized actions in the video history as prompts for Llama2-7B to predict future actions. Likewise, Palm uses the history of recognized actions and narrations from the video history as prompts to GPTNeo-1.3B [5] for future action prediction. **VPA baselines.** Akin to LTA, we provide a supervised non-LM (supervised Ego4D model [15, 33] trained on the VPA trainset) and a fine-tuned LLM-based approach (VLaMP [33]) as baselines for VPA. In addition to the Ego4D supervised baseline, which uses a SlowFast video encoder followed by classification heads, we also provide a random action prediction and an LSTM-based supervised baseline called

[3] We choose VLaMP over VideoLLM [7] as our finetuned LM baseline because of the availability of code for the former. VideoLLM also only provides results on LTA's test set, preventing a direct comparison on the v1 set.

| Model | Visual History | Visual Encoder | ED@(Z=5)↓ | | | ED@(Z=20)↓ | | |
|---|---|---|---|---|---|---|---|---|
| | | | Verb | Noun | Action | Verb | Noun | Action |
| AntGPT [48] | T | CLIP | - | - | - | 0.756 | 0.725 | - |
| Palm*[4] [16] | T | Blip2 | - | - | - | 0.732 | 0.812 | 0.958 |
| Socratic 13B | T | LaViLa | 0.689 | 0.681 | 0.919 | 0.731 | 0.732 | 0.929 |
| Socratic 70B | T | LaViLa | **0.683** | **0.661** | **0.917** | **0.726** | **0.712** | **0.928** |
| VCLM 13B | V+T | Internvideo+LaViLa | 0.698 | 0.685 | 0.925 | 0.740 | 0.751 | 0.932 |
| VCLM 70B | V+T | Internvideo+LaViLa | 0.696 | 0.669 | 0.923 | 0.739 | 0.731 | 0.931 |
| Ego4D (supervised) [15] | V | SlowFast | - | - | - | 0.745 | 0.779 | 0.941 |
| VLaMP (supervised) [33] | V+T | S3D | - | - | - | 0.730 | 0.772 | 0.932 |

Table 2. **Long-term action anticipation on Ego4D.** Edit distance values for forecasting horizon of $Z = 5$ and $Z = 20$ actions are shown on v1 validation set.

DDN [6] from the VPA paper [33].

### 4.3. Quantitative Results

**Text-based representation of visual history is more effective than implicit representation when encoding long visual history.** We find that the Socratic approach outperforms VCLMs for predicting actions in Ego4D LTA irrespective of the LLM size (Table 2). Our results suggest that the visual embeddings used by VCLMs are less amenable for encoding long visual histories. Specifically, the implicit information contained in these visual embeddings does not add much beyond the text-based representation of visual history for future prediction tasks that require grounding longer visual histories. This finding is consistent across different future prediction horizons as highlighted by $Z = 5$ and $Z = 20$ results. Recall that irrespective of the history lengths, our Socratic and VCLMs have a fixed context window. While Socratic models use the entire context to encode text, VCLMs use $12.5\%$ of their available tokens in the context window for visual embeddings. For long-history tasks, our results suggest that it is better to use the available context window to encode the history as text rather than devoting tokens to capture implicit information.

**Smaller LLMs benefit from implicit representation of visual information for short to medium-range visual history.** In contrast to LTA however, VCLMs show competitive performance in the VPA task requiring grounding of medium visual history (Table 3). Specifically, VCLM 13B outperforms the Socratic 13B model by a large margin (mAcc: $21.2 \rightarrow 25.5$ and mIoU: $37.4 \rightarrow 45.5$ for $Z = 4$) in VPA on CrossTask (Table 3). In Appendix 9.1, we show that this trend is consistent for 7B LLMs. However, this performance gap between the VCLMs and Socratic models disappears for 70B LLMs in VPA. Thus, the implicit vision representation may capture signals that help in forecasting, especially when using smaller LLMs (7B, 13B) with limited reasoning and planning capabilities. However, such implicit information may not be essential for larger LLMs, which may be able to plan well with coarse-level grounding.

**Larger language models lead to better planning with limited, unstructured information from the visual history.** Akin to many existing works [18], we find that scaling laws hold for our video-based planning tasks. Overall action prediction performance improves in both LTA and VPA for both categories of models as LLM size increases from 13B to 70B (Tables 2,3). Furthermore, in comparison to existing SOTA approaches such as AntGPT [48] (7B/13B) and Palm [16] (1.5B/7B) in Table 2, our Socratic model (70B) achieves better performance with limited and unstructured information from visual history. Specifically, AntGPT uses Llama2 13B to infer the overall goal of actions in visual history and leverages it with Chain of Thought (COT) prompting [43] to predict future actions. Palm uses an additional action recognition model – EgoVLP [25] to extract actions present in the visual history and uses them along with the narrations to represent the visual history. Instead, our models only use narrations to represent the visual history. Further scaling comparisons can be found in Appendix 9.2.

[4]We include Palm* as the reproduced version of Palm [16].

| Model | Supervised Samples | $Z = 1$ | $Z = 3$ | | | $Z = 4$ | | |
|---|---|---|---|---|---|---|---|---|
| | | mAcc | SR | mAcc | mIOU | SR | mAcc | mIOU |
| Random | | 0.9 | 0.0 | 0.9 | 1.5 | 0.0 | 0.9 | 1.9 |
| Socratic 13B | 8 | 22.8 | 5.6 | 22.2 | 35.6 | 3.0 | 21.2 | 37.4 |
| VCLM 13B | 8 | 27.2 | 6.9 | 25.2 | 41.7 | 4.3 | 25.5 | 45.5 |
| Socratic 70B | 8 | 28.1 | **9.1** | 26.6 | **43.6** | 5.5 | 25.5 | 45.7 |
| VCLM 70B | 8 | 28.1 | 8.9 | **26.9** | 43.4 | **6.1** | **26.8** | **46.9** |
| DDN (supervised) [6, 33] | 1756 | 33.4 | 6.8 | 25.8 | 35.2 | 3.6 | 24.1 | 37.0 |
| Ego4D (supervised) [15] | 1756 | 26.9 | 2.4 | 24.0 | 35.2 | 1.2 | 21.7 | 36.8 |
| VLaMP (supervised) [33] | 1756 | 50.6 | 10.3 | 35.3 | 44.0 | 4.4 | 31.7 | 43.4 |

Table 3. **Short and medium horizon visual planning (VPA) on CrossTask.** Mean accuracy, mean IoU, and Success Rate (SR) percentages are shown for short, $Z = 1$, and medium, $Z = 3, 4$, horizons. We follow the prior work on VPA [33] for generating narrations while Internvideo [42] is used as the video encoder for VCLM as in LTA.

**LLMs reduce the need for supervision in long-horizon planning.** Few-shot Socratic and VCLMs achieve performance comparable to the fully supervised models on LTA (Table 2) as well as on VPA for longer-horizon predictions $Z = 4$ (Table 3). Notably, as highlighted in Tab. 3, such competitive performance is achieved while using only a fraction of supervision (8 examples in few-shot models vs. 1756 examples for finetuning VLaMP). Overall, the ability of the models to capture a broader but probable distribution of action cooccurrences and dependencies given coarse context may be more important for longer horizon predictions. Consequently, the common sense of LLMs shines in these settings leading to competitive performance despite no task-specific finetuning.

## 5. User-in-the-loop Evaluation

Our benchmarking experiments on LTA and VPA highlight the strengths and weaknesses of VCLMs and Socratic approaches to predict future actions based on video history. However, it is unclear how these actions would manifest with a user in the loop and whether these actions – when executed – would successfully complete real-world activities. To that end, we conduct an online evaluation of VCLMs and Socratic approaches in real-world assistive scenarios. We recruit 18 participants to perform multi-step cooking activities while wearing an egocentric observation device called Aria [37] and following assistance from one of these models. We measure the true activity completion success rate, which is difficult to measure offline, and the correctness of recommended actions using mean IoU.

### 5.1. Study Design

**Multi-step activities.** We choose three cooking activities for our study: 1) espresso latte, 2) caprese salad,

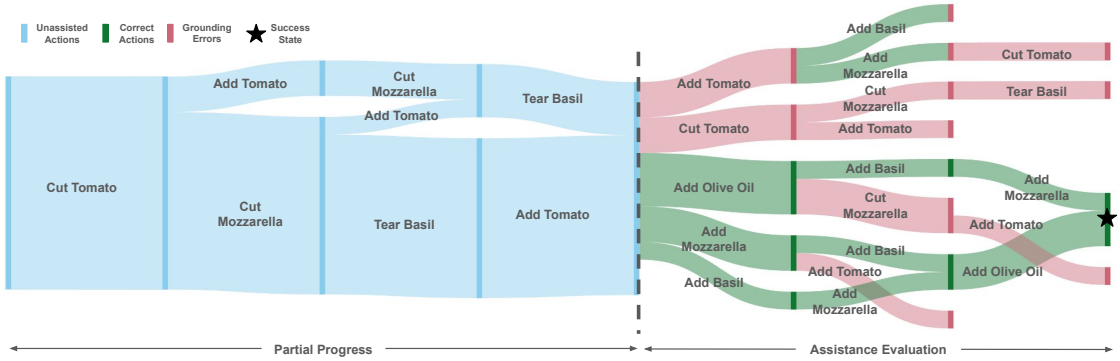



Figure 3. **Overview of the flow of steps for the activity of making caprese salad**. Unassisted actions performed by the participants in the *partial progress* phase of the activity followed by multimodal LLM assisted activity execution is shown.

and 3) BLT (bacon lettuce and tomato) sandwich. These activities consist of a variety of ingredients, including meat, vegetables, breads, and liquids, and require different types of actions, including pouring, chopping, spreading, and plating. Furthermore, these activities also have some ordering constraints among the steps. For instance, the milk needs to be frothed before pouring into the espresso for making latte and the BLT ingredients need to be stacked on the bread before closing the sandwich. Lastly, we account for the ease of doing these activities in an office kitchen, which required omitting really long activities or activities using a stove, oven, etc.

**Study protocol.** Each participant performs two of the three aforementioned activities instructed by either a VCLM or a Socratic model. We use Latin-square counterbalancing for the ordering of multimodal LLM that offers them assistance as well as the type of activity that they perform across participants to reduce learning effect [17].

Each activity entails a script that a participant can follow. Full scripts of the three activities can be found in Appendix 11. Each activity script is split into two phases – 1) *Partial progress:* In the initial phase of the activity, participants are asked to make partial progress in the activity by completing a set of steps in any feasible order. For example, in the caprese-making activity, participants are instructed to slice tomatoes and mozzarella, tear basil, and place tomato slices on the plate. They are also free to slice varying amounts of these ingredients in whatever manner they like e.g., small vs. large slices. 2) *Assistance evaluation:* In this phase of the activity, the multimodal LLM assistant engages and guides the participant through completing the remaining steps. Participants iteratively request the next task step from the assistant and then execute what the assistant asks them to do to the best of their ability. Figure 3 gives an overview of these phases for the caprese-making activity. Participants may skip recommended actions from the assistant that are infeasible, irrelevant, or already completed. Actions are skipped solely at the participant's discretion. The activity episode is considered complete if the assistant returns a "done" step (for example, asking the participant to serve their dish), if the participant chooses to skip 3 instructions in a row, or after the participant executes $n+2$ actions, where $n$ is the number of steps in the evaluation section of the script. We allow $n+2$ actions so as to account for multiple successful action sequences, including ones with optional steps (Fig. 4).

**Evaluation protocol.** At the end of each activity episode, participants are asked to evaluate whether the food item they produced with the assistant's help is consistent with their idea of the food item they were supposed to prepare in the activity. If they are unfamiliar with the food item, they may conduct an internet search first to determine the characteristics of the item. Independently, the study administrator also evaluates whether the participant's product matches the food description. We consider an activity episode to be successful if both the participant and the administrator rate the episode as successful. Since the activity can be accomplished in multiple ways and with optional steps, our approach of using two human ratings for estimating activity success ensures a conservative and robust measurement. We also record individual actions recommended by the assistant and whether they were skipped, executed, or infeasible to compute the mean IoU and analyze the types of errors.

## 5.2. Real-world Deployment of Multimodal LLMs

We frame the multimodal LLM assistance in our online study after the VPA task [33] used in our offline benchmarking experiments (Sec. 4), since we believe its definition is closest to vision-based assistants. Following VPA, the VCLMs and Socratic models are given an untrimmed and unsegmented egocentric video stream from the *partial progress* phase of the activity. This usually corresponds to 3-5 high-level actions on average. We also provide the models with a natural language goal describing the activity, as in VPA. The models are then prompted to iteratively output single-step action predictions to guide the user through the remaining 2-3 steps in the *assistance evaluation* phase of each activity. Akin to our offline benchmarking, the models in our study did not have access to the activity scripts, nor had they seen the kitchen environment where the experiments were conducted. We use the same retrieval-based few-shot prompting strategy as in our offline experiments to obtain predictions from these models.

**Model modifications for offline $\rightarrow$ online.** To keep inference times short during the study, we only use 13B versions of our VCLMs and Socratic models. However, direct de-

ployment of these offline models on the online video stream from Aria does not work out of the box. The visual history accumulated in *partial progress* phase of activities in the study can consist of up to 1500 frames, corresponding to 2+ minutes of video, and are akin to the visual histories in LTA. However, unlike LTA, where ground-truth segmentation of these long video histories is available to generate text-based representations and vision embeddings for Socratic and VCLM, respectively, the video history from Aria is unsegmented. To support the grounding of long, unsegmented visual history in Socratic and VCLM, we make two main modifications. First, we perform segmentation. However, the addition of yet another model, e.g., a video segmentation model in our processing pipeline, could increase computation time and lead to interaction delays in our user-in-the-loop setup. Therefore, we uniformly segment the Aria stream into clips before passing them to our narration model – LaViLa. To compensate for unrelated and repeated narrations emerging from the uniformly segmented stream, we generate and cluster multiple narrations per segment as well as across segments based on the semantic similarity of narrations. Despite such stream segmentation and narration clustering, we find that the narrations tend to be extremely low-level, which leads to a very long narration history – ultimately exceeding the context window of our multimodal LLMs. Hence, our second modification entails the addition of a goal-conditioned summarization step to produce the final set of narrations for encoding the long visual history in online settings. Appendix 7.2 provides additional details about these modifications. Lastly, unlike offline benchmarking experiments where we match the open-set model outputs to a closed-set of actions (Sec 4), we directly use the open-set output for easier interactions with the user in the loop. No other modifications were made to the models for online deployment. Table 5 in the appendix shows these online-modified models perform similarly on the VPA task.

**System setup.** We obtain the RGB video stream from Aria donned by our participants at 10 frames per second over wifi to a local machine. The frames are then center-cropped and downsampled in resolution ($1400 \times 1400 \rightarrow 288 \times 384$) to match the resolution of the LaViLa encoder. These frames are then sent to a remote server, which hosts the multimodal LLMs. The step suggestions returned by the models are parsed and communicated to the user via text-to-speech over wireless earbuds. The LaViLa narrator model runs on two-second clips of video and outputs 10 narrations per clip pre-clustering. The summarization step runs over the entire clustered narration history before every prediction step. Further details on online setup and inference can be found in Appendix 7.3.

### 5.3. Quantitative Results

**The Socratic appraoch outperforms VCLMs at user-in-the-loop activity assistance.** Table 4 shows the results of our online study. Akin to our offline experiments that showed the Socratic approach outperforming VCLMs in LTA (Table 2) and demonstrate competitive performance with VCLMs in VPA (Table 3), we find that the Socratic approach enables higher activity completion success rate as well as mIoU across the 18 participants and 3 activities in our online study. Note that our online Socratic model does not leverage finetuned video narration models like our offline experiments. Nevertheless, it exhibits superior performance. Despite the low overall success rate of both models, in 40% of successful trials, the Socratic model enabled a user to complete a task they had not previously done.

| Method | Visual History | Success Rate | mIoU |
|---|---|---|---|
| Socratic 13B | T | **27.8** | **30.4** |
| VCLM 13B | V+T | 16.7 | 23.0 |
| Socratic 13B (Offline) | T | - | 40.3 |
| VCLM 13B (Offline) | V+T | - | **44.0** |

Table 4. **Activity completion success rate and mean IoU metrics in percentage for the online study across all participants and activities.** Socratic models outperform VCLMs in online settings. We also rerun the models offline on the videos collected from our study to compare offline vs. online metrics. While success rate cannot be compared across offline and online settings, we find that mIoU also isn't consistent between the two settings. Appendix 12 contains a detailed breakdown of failure modes in online experiments.

**Offline metrics do not capture online performance.** The success rate of activity completion cannot be truly measured in offline datasets. However, it is unclear whether other metrics used for evaluating video-based action anticipation and planning such as mIoU and edit distance (Sec. 4) translate from offline settings to online settings. Despite being small scale as compared to datasets in our offline benchmarking experiments, the video data from our online experiments enable a unique opportunity to compare these metrics head-on in both online and offline settings. To this end, we rerun Socratic and VCLM offline in the videos from our study. Specifically, we provide the models with videos from the *partial progress* phase of the activities along with the activity goal, e.g., "make Caprese salad with mozzarella, tomato, basil, olive oil" following the VPA task (Sec. 4). The models are then prompted with few-shot examples, following our prompt setup from offline experiments, to predict $n+2$ steps. Here, $n$ is the expected number of steps remaining in the activity from the *assistance evaluation* phase of the activity. We observe higher mean IoU rates for both models when run offline compared to their mIoU when they provided user-in-the-loop assistance online. Furthermore, VCLM outperforms Socratic in offline mIoU. However, it lags behind in both online mIoU and real-world success rates, indicating that offline mIoU may be an unreliable predictor of real-world performance. The gap between offline and online mIoU may partly be attributed to the single multi-step prediction of all the remaining actions in the offline setting versus iterative single-step predictions, i.e., with replanning in the online setting. The iterative single-step predictions are more likely to make repeat suggestions, often due to grounding errors, which lead to a lower intersection between suggested steps and ground-truth steps.

### 5.4. Qualitative Analysis of Model Errors

**Grounding errors, planning errors, and failure to detect activity end/success are the main failure modes.** We also evaluate cases where a participant skipped actions recom-

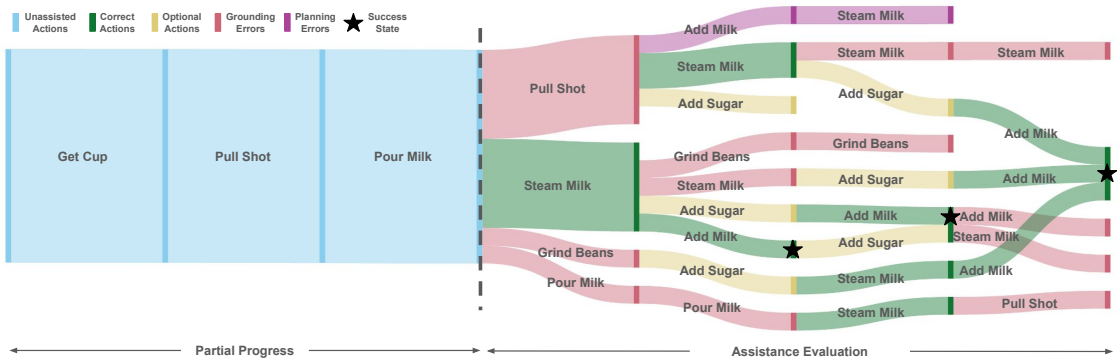



Figure 4. **Various error modes of multimodal LLMs in the latte activity**. Models fail to ground the steps that are already completed, recommend steps with incorrect ordering (planning error), or fail to recognize that the activity is complete.

mended by the assistant. Recall that participants could skip action recommendations that were redundant, infeasible, or irrelevant. Appendix 12 shows a detailed analysis and breakdown of these reasons for skipped actions by the participants across the VCLM and Socratic models per activity. Such analysis enabled us to identify three main error modes – grounding errors, planning errors, and failure to detect activity end/success. Figure 4 shows these error modes for the espresso latte activity across all participants. Redundant skipped actions often correlate to grounding mistakes, where the models fail to recognize a step that has already been completed. Infeasible skipped steps also often correlate with grounding errors. Here, the models may suggest something that works for a different variation of the activity. For example, grinding coffee might work for a different version of a latte-making activity, but participants used an automated espresso machine in our study. These grounding errors are often more subtle than grounding errors from redundant steps. Finally, irrelevant skipped steps often correlate to planning errors, where the models suggest a step that is not part of the activity. We also find that in $50\%$ of the successful activity episodes, the models fail to recognize when an activity is completed.

**Offline metrics don't capture error modes.** The failure of models to detect activity end/success state does not affect the success rate metric, which would count such activity episodes as successful. However, it does lower mIoU scores due to redundant suggested actions. The overview of the participant steps for making a latte in Fig. 4 succinctly highlights the known issues with offline metrics. Specifically, mIoU as a permissive metric, would consider adding milk before steaming, a planning error, as a success. Conversely, offline success rate, being a restrictive metric, would discount 4 of the 5 present paths to success for making lattes as failures. Furthermore, mIoU and edit distance metrics do not capture optional actions sometimes suggested by such models that do not affect activity success e.g., adding sugar.

**Grounding errors are the dominant mode of failure for models online.** The bulk of the errors both models exhibit pertain to grounding. In particular, past participant actions are either not captured by the narrations or visual embedding of their activity history or are present in the long history but not attended to by the models during prediction, leading to grounding errors. We find that 63% of the skipped action recommendations are due to redundant action suggestions emerging from erroneous grounding (Appendix 12). The distribution of skipped actions is consistent across both models and indicates that recognizing previously completed actions in an activity is an ongoing challenge for these models. In contrast, both models make fewer planning errors, i.e., they suggest fewer irrelevant actions or actions with incorrect orderings. Overall, our analyses of skipped actions and errors in the study indicate that the primary challenge with visual assistants still lies in reasoning about activity progress and activity success/failure via grounding – more so than task knowledge or planning.

## 6. Conclusion

We evaluate the two predominant multimodal LLM-based approaches: Vision Conditioned Language Models (VCLM) and Socratic Models for vision-based activity assistance through two video-based action anticipation benchmarks on *offline* datasets and a real-world *online* study with 18 participants. To the best of our knowledge, our online evaluation is the first of its kind for multimodal LLMs towards real-world activity assistance systems. Our experiments show the Socratic approach is better equipped to capture coarse visual details across a long visual history. Current VCLMs can capture more fine-grained details but only for short visual history. Encoding long videos and aligning long videos with text tokens as needed by VCLMs would thus be rich avenues for future work. In the interim, Socratic models demonstrate competitive behaviors on video-based action anticipation and planning tasks spanning short to long visual history both offline and online.

Our work sets important directions for future research on multimodal LLMs as vision-based assistants. Our online study demonstrates grounding is the largest source of errors for these models. Grounding at different granularities remains an open problem, which, when improved, will greatly enhance activity assistance systems. Furthermore, we highlight the performance gap between offline and online success, demonstrating the importance of real-world evaluation of models for assistive scenarios.

## References


[1] Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in Neural Information Processing Systems **35**, 23716–23736 (2022) 2, 3, 16

[2] Anderson, P., Wu, Q., Teney, D., Bruce, J., Johnson, M., Sünderhauf, N., Reid, I., Gould, S., van den Hengel, A.: Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3674–3683 (2018) 2

[3] Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., Parikh, D.: VQA: Visual question answering. In: ICCV (2015) 2

[4] Awadalla, A., Gao, I., Gardner, J., Hessel, J., Hanafy, Y., Zhu, W., Marathe, K., Bitton, Y., Gadre, S., Sagawa, S., et al.: Openflamingo: An open-source framework for training large autoregressive vision-language models. arXiv preprint arXiv:2308.01390 (2023) 2

[5] Black, S., Leo, G., Wang, P., Leahy, C., Biderman, S.: GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow (Mar 2021). https://doi.org/10.5281/zenodo.5297715, https://doi.org/10.5281/zenodo.5297715, If you use this software, please cite it using these metadata. 4

[6] Chang, C.Y., Huang, D.A., Xu, D., Adeli, E., Fei-Fei, L., Niebles, J.C.: Procedure planning in instructional videos. In: European Conference on Computer Vision. pp. 334–350. Springer (2020) 5

[7] Chen, G., Zheng, Y.D., Wang, J., Xu, J., Huang, Y., Pan, J., Wang, Y., Wang, Y., Qiao, Y., Lu, T., et al.: Videollm: Modeling video sequence with large language models. arXiv preprint arXiv:2305.13292 (2023) 1, 2, 4

[8] Cheng, Z., Leng, S., Zhang, H., Xin, Y., Li, X., Chen, G., Zhu, Y., Zhang, W., Luo, Z., Zhao, D., Bing, L.: Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms (2024), https://arxiv.org/abs/2406.07476 1, 2, 3, 4, 17

[9] Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning (2023) 2

[10] Damen, D., Fragomeni, A., Munro, J., Perrett, T., Whettam, D., Wray, M., Furnari, A., Farinella, G.M., Moltisanti, D.: Epic-kitchens-100-2021 challenges report (2021) 1, 3

[11] Das, A., Kottur, S., Gupta, K., Singh, A., Yadav, D., Moura, J.M., Parikh, D., Batra, D.: Visual dialog. In: CVPR (2017) 2

[12] Driess, D., Xia, F., Sajjadi, M.S.M., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., Huang, W., Chebotar, Y., Sermanet, P., Duckworth, D., Levine, S., Vanhoucke, V., Hausman, K., Toussaint, M., Greff, K., Zeng, A., Mordatch, I., Florence, P.: Palm-e: An embodied multimodal language model. In: arXiv preprint arXiv:2303.03378 (2023) 2

[13] Gao, P., Han, J., Zhang, R., Lin, Z., Geng, S., Zhou, A., Zhang, W., Lu, P., He, C., Yue, X., Li, H., Qiao, Y.: Llama-adapter v2: Parameter-efficient visual instruction model. arXiv preprint arXiv:2304.15010 (2023) 2

[14] Gong, T., Lyu, C., Zhang, S., Wang, Y., Zheng, M., Zhao, Q., Liu, K., Zhang, W., Luo, P., Chen, K.: Multimodal-gpt: A vision and language model for dialogue with humans. arXiv preprint arXiv:2305.04790 (2023) 2

[15] Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., Martin, M., Nagarajan, T., Radosavovic, I., Ramakrishnan, S.K., Ryan, F., Sharma, J., Wray, M., Xu, M., Xu, E.Z., Zhao, C., Bansal, S., Batra, D., Cartillier, V., Crane, S., Do, T., Doulaty, M., Erapalli, A., Feichtenhofer, C., Fragomeni, A., Fu, Q., Fuegen, C., Gebreselasie, A., Gonzalez, C., Hillis, J., Huang, X., Huang, Y., Jia, W., Khoo, W., Kolar, J., Kottur, S., Kumar, A., Landini, F., Li, C., Li, Y., Li, Z., Mangalam, K., Modhugu, R., Munro, J., Murrell, T., Nishiyasu, T., Price, W., Puentes, P.R., Ramazanova, M., Sari, L., Somasundaram, K., Southerland, A., Sugano, Y., Tao, R., Vo, M., Wang, Y., Wu, X., Yagi, T., Zhu, Y., Arbelaez, P., Crandall, D., Damen, D., Farinella, G.M., Ghanem, B., Ithapu, V.K., Jawahar, C.V., Joo, H., Kitani, K., Li, H., Newcombe, R., Oliva, A., Park, H.S., Rehg, J.M., Sato, Y., Shi, J., Shou, M.Z., Torralba, A., Torresani, L., Yan, M., Malik, J.: Ego4d: Around the World in 3,000 Hours of Egocentric Video. In: IEEE/CVF Computer Vision and Pattern Recognition (CVPR) (2022) 1, 2, 3, 4, 5, 12

[16] Huang, D., Hilliges, O., Van Gool, L., Wang, X.: Palm: Predicting actions through language models@ ego4d long-term action anticipation challenge 2023. arXiv preprint arXiv:2306.16545 (2023) 1, 2, 3, 4, 5, 14, 17

[17] John, P.W.: Statistical design and analysis of experiments. SIAM (1998) 6

[18] Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., Amodei, D.: Scaling laws for neural language models. arXiv preprint arXiv:2001.08361 (2020) 5

[19] Laurençon, H., Saulnier, L., Tronchon, L., Bekman, S., Singh, A., Lozhkov, A., Wang, T., Karamcheti, S., Rush, A.M., Kiela, D., Cord, M., Sanh, V.: Obelics: An open web-scale filtered dataset of interleaved image-text documents (2023) 2

[20] Li, B., Zhang, Y., Chen, L., Wang, J., Yang, J., Liu, Z.: Otter: A multi-modal model with in-context instruction tuning. arXiv preprint arXiv:2305.03726 (2023) 2

[21] Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 (2023) 2, 14

[22] Li, K., He, Y., Wang, Y., Li, Y., Wang, W., Luo, P., Wang, Y., Wang, L., Qiao, Y.: Videochat: Chat-centric video understanding. arXiv preprint arXiv:2305.06355 (2023) 1, 2, 4, 17

[23] Li, L., Yin, Y., Li, S., Chen, L., Wang, P., Ren, S., Li, M., Yang, Y., Xu, J., Sun, X., et al.: Mit: A large-scale dataset towards multi-modal multilingual instruction tuning. arXiv preprint arXiv:2306.04387 (2023) 2

[24] Lin, B., Ye, Y., Zhu, B., Cui, J., Ning, M., Jin, P., Yuan, L.: Video-llava: Learning united visual representation by alignment before projection (2023), https://arxiv.org/abs/2311.10122 1, 2, 3, 4, 16, 17

[25] Lin, K.Q., Wang, A.J., Soldan, M., Wray, M., Yan, R., Xu, E.Z., Gao, D., Tu, R., Zhao, W., Kong, W., et al.: Egocentric video-language pretraining. arXiv preprint arXiv:2206.01670 (2022) 5

[26] Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., Lee, Y.J.: Llava-next: Improved reasoning, ocr, and world knowledge (2024) 1, 2, 3, 4, 17

[27] Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning (2023) 2, 3

[28] Lyu, C., Wu, M., Wang, L., Huang, X., Liu, B., Du, Z., Shi, S., Tu, Z.: Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration. arXiv preprint arXiv:2306.09093 (2023) 2

[29] Maaz, M., Rasheed, H., Khan, S., Khan, F.S.: Video-chatgpt: Towards detailed video understanding via large vision and language models. arXiv preprint arXiv:2306.05424 (2023) 1, 2, 17

[30] Miech, A., Zhukov, D., Alayrac, J.B., Tapaswi, M., Laptev, I., Sivic, J.: Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2630–2640 (2019) 16

[31] Moon, S., Madotto, A., Lin, Z., Nagarajan, T., Smith, M., Jain, S., Yeh, C.F., Murugesan, P., Heidari, P., Liu, Y., et al.: Anymal: An efficient and scalable any-modality augmented language model. arXiv preprint arXiv:2309.16058 (2023) 1, 2, 3, 4, 16, 17

[32] Nagarajan, T., Li, Y., Feichtenhofer, C., Grauman, K.: Ego-topo: Environment affordances from egocentric video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 163–172 (2020) 3

[33] Patel, D., Eghbalzadeh, H., Kamra, N., Iuzzolino, M.L., Jain, U., Desai, R.: Pretrained language models as visual planners for human assistance. arXiv preprint arXiv:2304.09179 (2023) 1, 2, 3, 4, 5, 6, 12, 13, 14, 15

[34] Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al.: Improving language understanding by generative pre-training (2018) 2

[35] Reed, S., Zolna, K., Parisotto, E., Colmenarejo, S.G., Novikov, A., Barth-Maron, G., Gimenez, M., Sulsky, Y., Kay, J., Springenberg, J.T., et al.: A generalist agent. arXiv preprint arXiv:2205.06175 (2022) 2

[36] Ren, S., Yao, L., Li, S., Sun, X., Hou, L.: Timechat: A time-sensitive multimodal large language model for long video understanding (2024), https://arxiv.org/abs/2312.02051 17

[37] Somasundaram, K., Dong, J., Tang, H., Straub, J., Yan, M., Goesele, M., Engel, J.J., De Nardi, R., Newcombe, R.: Project aria: A new tool for egocentric multi-modal ai research. arXiv preprint arXiv:2308.13561 (2023) 1, 2, 5

[38] Su, Y., Lan, T., Li, H., Xu, J., Wang, Y., Cai, D.: Pandagpt: One model to instruction-follow them all. arXiv preprint arXiv:2305.16355 (2023) 2

[39] Thakur, N., Reimers, N., Daxenberger, J., Gurevych, I.: Augmented SBERT: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 296–310. Association for Computational Linguistics, Online (Jun 2021), https://www.aclweb.org/anthology/2021.naacl-main.28 4

[40] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023) 3

[41] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023) 3, 4

[42] Wang, Y., Li, K., Li, Y., He, Y., Huang, B., Zhao, Z., Zhang, H., Xu, J., Liu, Y., Wang, Z., et al.: Internvideo: General video foundation models via generative and discriminative learning. arXiv preprint arXiv:2212.03191 (2022) 4, 5, 16

[43] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems **35**, 24824–24837 (2022) 5

[44] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: International conference on machine learning. pp. 2048–2057. PMLR (2015) 2

[45] Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., Wang, J., Hu, A., Shi, P., Shi, Y., et al.: mplug-owl: Modularization empowers large language models with multimodality. arXiv preprint arXiv:2304.14178 (2023) 2

[46] Zeng, A., Wong, A., Welker, S., Choromanski, K., Tombari, F., Purohit, A., Ryoo, M., Sindhwani, V., Lee, J., Vanhoucke, V., et al.: Socratic models: Composing zero-shot multimodal reasoning with language. arXiv preprint arXiv:2204.00598 (2022) 1, 2, 3

[47] Zhang, H., Li, X., Bing, L.: Video-llama: An instruction-tuned audio-visual language model for video understanding. arXiv preprint arXiv:2306.02858 (2023) 1, 2, 17

[48] Zhao, Q., Zhang, C., Wang, S., Fu, C., Agarwal, N., Lee, K., Sun, C.: Antgpt: Can large language models help long-term action anticipation from videos? arXiv preprint arXiv:2307.16368 (2023) 1, 2, 4, 5, 14, 17

[49] Zhao, Y., Misra, I., Krähenbühl, P., Girdhar, R.: Learning video representations from large language models. In: CVPR (2023) 4, 13, 14, 15, 16

[50] Zhou, X., Girdhar, R., Joulin, A., Krahenbuhl, P., Misra, I.: Detecting twenty-thousand classes using image-level supervision. ArXiv **abs/2201.02605** (2022), https://api.semanticscholar.org/CorpusID:245827815 14, 16

[51] Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023) 2

[52] Zhukov, D., Alayrac, J.B., Cinbis, R.G., Fouhey, D., Laptev, I., Sivic, J.: Cross-task weakly supervised learning from instructional videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3537–3545 (2019) 3, 12