Socratic Models: Composing Zero-Shot Multimodal Reasoning with Language

Andy Zeng, Adrian Wong, Stefan Welker, Krzysztof Choromanski, Federico Tombari,
Aveek Purohit, Michael Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, Pete Florence
Google

Abstract

Large foundation models can exhibit unique capabilities depending on the domain of data they are trained on. While these domains are generic, they may only barely overlap. For example, visual-language models (VLMs) are trained on Internet-scale image captions, but large language models (LMs) are further trained on Internet-scale text with no images (e.g. from spreadsheets, to SAT questions). As a result, these models store different forms of commonsense knowledge across different domains. In this work, we show that this model diversity is symbiotic, and can be leveraged to build AI systems with structured Socratic dialogue – in which new multimodal tasks are formulated as a guided language-based exchange between different pre-existing foundation models, without additional finetuning. In the context of egocentric perception, we present a case study of Socratic Models (SMs) that can provide meaningful results for complex tasks such as generating free-form answers to contextual questions about egocentric video, by formulating video Q&A as short story Q&A, i.e. summarizing the video into a short story, then answering questions about it. Additionally, SMs can generate captions for Internet images, and are competitive with state-of-the-art on zero-shot video-to-text retrieval with 42.8 R@1 on MSR-VTT 1k-A. SMs demonstrate how to compose foundation models zero-shot to capture new multimodal functionalities, without domain-specific data collection. Prototypes are available at socraticmodels.github.io.

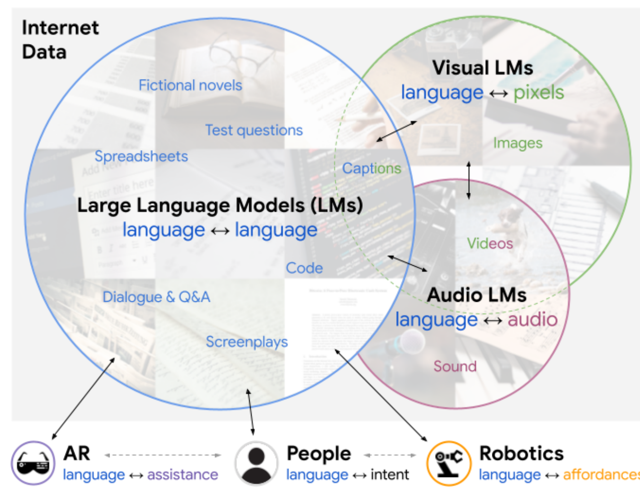


Fig. 1: Language-based foundation models trained across different domains learn complementary forms of commonsense. Language is an intermediate representation by which these models can communicate with each other to generate joint predictions for new multimodal tasks, without finetuning. New applications (e.g., augmented reality (AR), human feedback, robotics) can be viewed as adding participants to the multi-model discussion. In this paper, we study inter-model dialogue in the context of multimodal reasoning.

1 Introduction

Foundation models (Bommasani et al. 2021) (e.g., BERT, GPT-3, CLIP) have enabled impressive capabilities in recent years: from zero-shot image classification (Radford et al. 2021; Li et al. 2021a), to high-level planning (Huang et al. 2022; Ahn et al. 2022). These capabilities depend on their training data distribution – and while they may be generic or indiscriminately crawled from the web, their distributions remain distinct across domains. For example, in terms of linguistic data, visual-language models (VLMs) (Wang et al. 2021; Jain et al. 2021) are trained on image and video captions, but large language models (LMs) (Devlin et al. 2018; Thoppilan et al. 2022; Chen et al. 2021) are additionally trained on a large corpora of other data such as spreadsheets, fictional novels, and standardized test questions. These different domains offer distinct commonsense knowledge: VLMs can ground text to visual content, but LMs can perform a variety of other linguistic tasks (e.g., answering reading comprehension questions (Rajpurkar, Jia, and Liang 2018)) that to date have not been demonstrated with VLMs alone. In this work, we propose that these model differences are complementary, and can be jointly leveraged to build AI systems with structured Socratic dialogue – in which new multimodal tasks are formulated as a guided exchange between different pre-existing language-based foundation models, without additional finetuning. Rather than scaling training data in the areas of overlap (e.g., alt-text captions (Jia et al. 2021)), or unifying model architectures (Hu and Singh 2021), Socratic Models[1] (SMs) are a class of systems that embrace the zero-shot capabilities of foundation models by engineering guided discussions between the independent models to reach a shared consensus on a task-specific output. SMs use language as the representation by which inter-domain foundation models can jointly be used for inference.

[1]The name Socratic Models draws inspiration from an analogy to the Socratic Method, but with language-interactable models that may produce or interpret language.

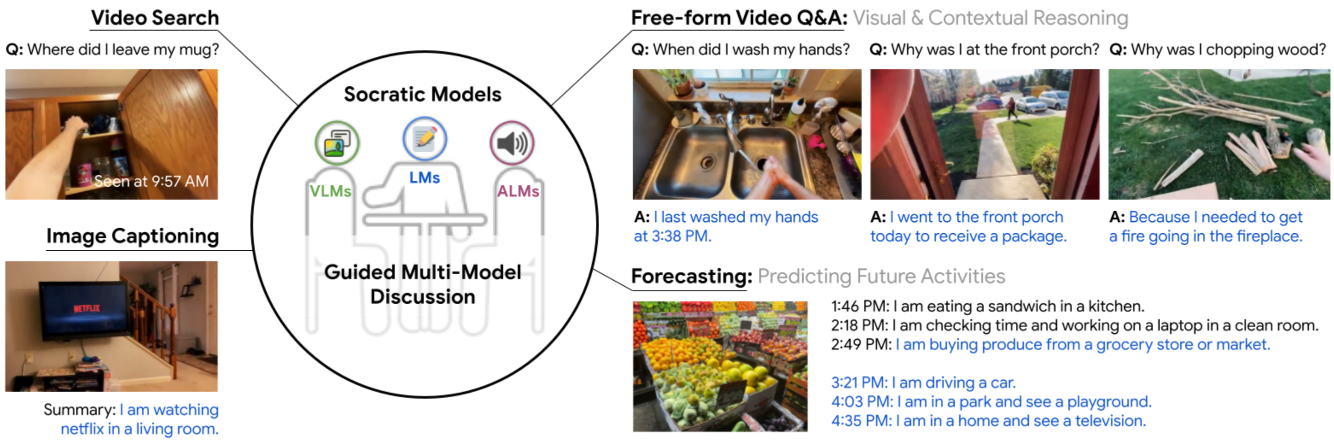


Fig. 2: In this work we propose Socratic Models (SMs), a framework that uses structured dialogue between pre-existing foundation models, each of which can exhibit unique (but complementary) capabilities depending on the distributions of data on which they are trained. On various perceptual tasks (shown), this work presents a case study of SMs with visual language models (VLMs, e.g., CLIP), large language models (LMs, e.g., GPT-3, RoBERTa), and audio language models (ALMs, e.g., Wav2CLIP, Speech2Text). From video search, to image captioning; from generating free-form answers to contextual reasoning questions, to forecasting future activities – SMs can provide meaningful results for complex tasks across classically challenging computer vision domains, without any model finetuning.

Across a number of tasks spanning vision, language, and audio modalities, we find that specific instantiations of SMs, using LMs together with VLMs and audio-language models (ALMs), can generate results on challenging perceptual tasks (examples in Fig. 2) that are often coherent and correct. We present results on Internet image captioning (Sec. 4) and the common video understanding task of video-to-text retrieval (Sec. 5), but our highlighted application is open-ended reasoning in the context of egocentric perception (Fig. 4) – from answering free-form contextual reasoning questions about first-person videos (e.g. *"why did I go to the front porch today?"*), to forecasting events into the future with commonsense (e.g. *"what will I do 3 hours from now?"*). Our egocentric SM system consists of two primary components, each of which benefits from multimodal multi-model discussions: (i) assembling video into a *language-based world-state history*, i.e. a story or event log, then (ii) performing various types of open-ended text-prompted tasks based on that world-state history. We find that simple scripted policies to guide a closed-loop exchange between pre-trained LM, VLM, and ALM models can (a) generate meaningful captions that respond to questions like *"what am I doing?"* with answers like *"receiving a package"* that span beyond the label set of standard vision datasets (Sigurdsson et al. 2018; Smaira et al. 2020), and (b) exhibit open-ended contextual Q&A capabilities previously thought to be out-of-reach for egocentric perception without domain-specific data collection (Grauman et al. 2021; Damen et al. 2020).

The goal of this paper is (1) to discuss new perspectives on building AI systems that embrace the heterogeneity of language-based foundation models through structured Socratic dialogue, and (2) give example demonstrations of what is already possible today with SMs on challenging perceptual tasks. Specifically, our contributions include (i) the Socratic Models framework, (ii) demonstration of an egocentric perception system using Socratic Models, (iii) qualitative results on video understanding (synthesizing video snippets from a full day of activity) that is not covered by existing benchmark datasets, (iv) qualitative comparisons to a state-of-the-art model (Mokady, Hertz, and Bermano 2021) on the task of single-image captioning in egocentric and Internet image domains, (v) quantitative comparisons to state-of-the-art video understanding models on the popular MSR-VTT (Xu et al. 2016; Yu, Kim, and Kim 2018) dataset for video-to-text retrieval, and (vi) a framework for unsupervised quantitative model selection of Socratic Models through sub-model ablations.

Overall, these ideas shed light on promising new opportunities to build simple systems for general applications that compose foundation models out-of-the-box. By construction, multimodal foundation models are likely to be trained on different distributions of Internet data, resulting in different test time capabilities. These capabilities can be improved for a given target distribution through finetuning, but at the cost of generality and robustness to distribution shifts (Wortsman et al. 2021). SMs offer an alternative approach in which these capabilities can be integrated cooperatively, and in which we can make use of concepts in domain A that are more easily obtained from domain B without complex alignment of the representations, i.e., through additional large large-scale training across multiple domains. Instead, the common representation is language, and it may be used to compose existing models, in a zero-shot manner.

Of course, our demonstrated SM systems are not without their limitations. We discuss these limitations, such as unreliability inherited from the foundation models on which they are constructed, together with other potential broader impacts (Sec. 8.1).

2 Socratic Models Framework

Socratic Models (SMs) is a framework (Fig. 2) in which multiple pretrained language-interactable foundation models are composed (Fig. 3) zero-shot to perform new downstream tasks e.g., egocentric perception (our primary highlight in this work), as well as Internet data image captioning and video understanding (i.e., video-to-text retrieval). While a common trend in multimodal learning has been to seek embedding spaces in which multiple modalities may co-exist, we instead compose multiple models with language. While we may leverage submodels built with shared embedding spaces, language is the representation itself upon which multiple of these models, as well as other language-interactable models, may interact. In closed-loop, multi-model interactions can perform joint inference. These guided multi-model exchanges are best described through examples, as provided in Sections 3.3, 3.4, 4, and 5.

This work explores the use of several classes of models, each capturing a different domain or functionality. For visualization purposes, outputs from LMs are blue, VLMs green, ALMs purple, other prompt text is gray, LM completions on a shown prompt are **bolded in blue**, and user inputs are magenta.

Large Language Models (LMs) at the scale of 100B or more parameters trained on Internet-scale data e.g., GPT-3 (Brown et al. 2020) have shown to achieve strong zero-shot performance on many NLP tasks including dialogue (Thoppilan et al. 2022), coding (Chen et al. 2021), and reasoning (Wei et al. 2022). LMs can generate language sequences autoregressively, or compute similarity measures between sentence embeddings (Devlin et al. 2018). In the SMs framework, LMs (and the knowledge that they store) are used to perform contextual reasoning and Q&A, conditioned on the information provided by other models. This reasoning may in turn be subsequently checked by other models as well.

Visual-Language Models (VLMs) such (Radford et al. 2021; Jia et al. 2021; Li et al. 2021a; Jain et al. 2021; Zhai et al. 2021) learn models that map images and text into a shared feature embedding space. VLMs can serve as image-text critics that can rank multiple text suggestions (either generic visual entity categories from existing datasets, or outputs from the LM) against a given image using cosine similarity between features. Functionally, this operates as an open-set vocabulary image classifier, which can be replaced with domain-specific alternatives e.g., (He et al. 2017). VLMs are often trained on filtered image alt-text HTML pairs crawled from the Internet (Sharma et al. 2018). In the SMs framework, VLMs are used to detect visual entities, across various categories, from large pre-existing dictionaries, and these detections can both inform and check the work of LMs used for reasoning. (Note: we indicate a VLM checking an LM's work, for example through ranking, with green-underlined blue text.)

Audio-Language Models (ALMs) ground audio information into language, and can take the form of speech-to-text models, or audio-text critics (Wu et al. 2021a; Zhao et al. 2021) that learn a semantic similarity measure between audio and text. We use both types of models to capture information from the audio domain (e.g., to rank generic sound categories such as "running water" against audio snippets, or to directly transcribe speech from audio inputs). (Note: we indicate a VLM checking an LM's work, for example through ranking, with purple-underlined blue text.)

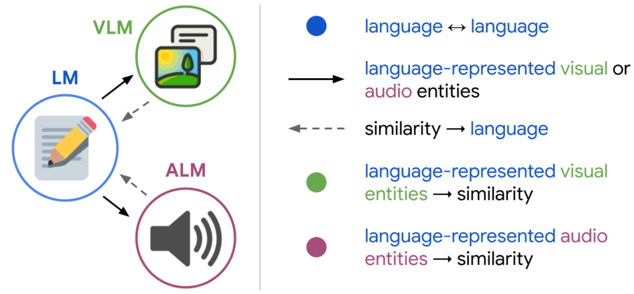


Fig. 3: Socratic Models are composed via guided multi-model exchanges. In this example, LMs, VLMs, and ALMs are composed closed-loop via language. Examples of guided multi-model discussions are provided in Sec. 3.3, Sec. 3.4, Sec. 4, and Sec. 5 on various multimodal applications.

Composing Multimodal Models. There are many options of foundation models to choose from, but the examples we showcase in the following sections use models that are publicly available, so that our SM systems can be made accessible to the community. In particular, we use CLIP (Radford et al. 2021) as the VLM (for egocentric perception, ViT-L/14 with 428M params); Wav2CLIP (Wu et al. 2021a) as the sound-critic ALM and Google Cloud Speech-to-text API as the speech-to-text ALM; and GPT-3 with 175B params (Brown et al. 2020; Ouyang et al. 2022) and RoBERTa (Liu et al. 2019b) with 355M params as the LMs. All pretrained models are used off-the-shelf with no additional finetuning.

In all our example systems shown, the composed foundation models are used zero-shot, and the interactions between these models are scripted. While in future work we are excited to explore the possibility of learning the interactions themselves, we also have found that there are practical benefits of harnessing a framework in which zero application-specific learning is performed. In particular, new applications can be quickly retargeted by a small amount of creative programming, with no training time or compute resources required. Additionally, the zero-shot capabilities may especially benefit domains in which training data is expensive to acquire (e.g., egocentric perception for AR, or robotics). The goal of this work is to demonstrate the capabilities available today in multimodal domains with Socratic Models.

In the following sections, we showcase example SM systems and results for egocentric perception (Sec. 3), generic Internet data image captioning (Sec. 4), and Internet data video-to-text retrieval (Sec. 5).

3 Socratic Models for Egocentric Perception

Here we describe an SM system for egocentric perception. We first provide a brief system overview (Section 3.1) and address "why egocentric perception?" (Section 3.2), then describe and show results for two primary system components: (i) compiling language-based world-state history (Section 3.3), and (ii) performing open-ended reasoning and video search on egocentric video (Section 3.4). We also explore methods to scale up video search capabilities (Section 3.5).

3.1 System Overview: Socratic Egocentric Perception

Overall, our system formulates egocentric video understanding as SM-guided discussions between different foundation models spanning multiple modalities. In general, our specific system can be thought of as a case study in the ways in which Socratic Models can be composed. Further, a goal of our egocentric perception system is to demonstrate progress towards performing unconstrained video understanding of egocentric videos. The *input* to this system is a video and an interactive sequence of natural language prompts (dialogue) from a user. For each prompt, depending on its type, the *output* is either a textual answer, and/or relevant frame(s) or video snippets from the video.

To this end, a key component of our system is using Socratic dialogue (Fig. 5) to turn key moments from an egocentric video into a language-based summary, which we term a *language-based world-state history* (Fig. 4, middle). This world-state history can then aid in prompting an LM to perform various reasoning tasks (Fig. 4, bottom) via text completion. In contrast to common paradigms for video Q&A, which may involve supervising video-to-text models on labeled datasets or with adversarial training (we refer to (Patel, Parikh, and Shastri 2021) for a recent survey), our approach simply maintains a language-based reconstruction of the video, generated by independently captioning key video frames, then recursively summarizing (Wu et al. 2021b) them to fit as input context to an LM (example in Fig. 5).

For open-ended reasoning, a key aspect of our approach is formulating *video understanding as reading comprehension*, i.e. re-frame "video Q&A" as a "short story Q&A" problem, where the language-based world-state history can be viewed as an interpretable state representation in the form of short stories, which can be used for reading comprehension Q&A, for which LMs have demonstrated strong zero-shot performance (Brown et al. 2020). Drawing analogies to 3D vision and robotics, this can be thought of as building an on-the-fly reconstruction of the sequence of events in the observable world with language, rather than other representations, such as dynamically-updated 3D meshes (Izadi et al. 2011) or neural fields (Tancik et al. 2022). In addition to open-ended question-answering in the form of text, we can also provide video search capabilities (in the form of image or audio retrieval, Fig 7) from natural language questions through zero-shot composition of SMs as well.

3.2 Why Egocentric Perception?

We highlight SMs on egocentric perception because it is an important yet challenging computer vision domain (Grauman et al. 2021; Damen et al. 2020; Sigurdsson et al. 2018) with downstream applications in augmented reality (AR) and robotics (Ahn et al. 2022). From unusual viewpoints to the lack of temporal curation – the characteristics of first-person videos are unique and not often found in existing datasets, which focus more on generic Internet content captured from third-person spectator views (Deng et al. 2009; Lin et al. 2014; Sharma et al. 2018). Notably, this domain shift makes it difficult for data-driven egocentric models to benefit from the standard paradigm of pretraining on third person Internet data (Li et al. 2021b; Sigurdsson et al. 2018). See Related Work (Sec. 7) for a more detailed discussion on prior work in egocentric perception – overall, the key challenges for the field have included how to acquire sufficient egocentric data, and/or how to make sufficient use of this data (either with dense labels, or otherwise).

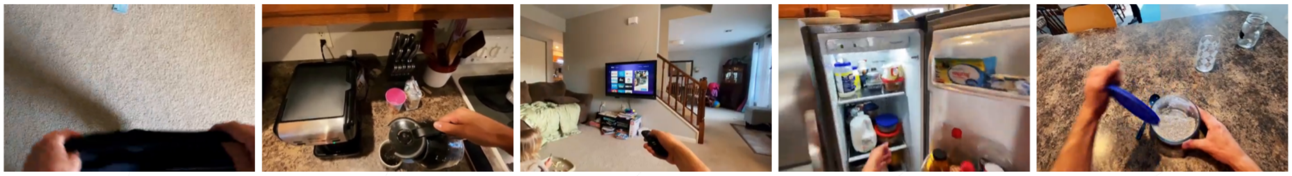
Despite the challenges of egocentric perception, we find that SMs can reconcile the complementary strengths of pretrained foundation models to address these difficulties through contextual reasoning. For example, while modern activity recognition models trained on third person data might over-index to the motion of the primary person in video (making the models difficult to be adapted to first-person videos), we find that LMs like GPT-3 can suggest equally plausible activities (e.g., "receiving a package") that may be occurring given only a brief description of the scene (e.g., "front porch") and the objects detected in the image ("package, driveway, door") by a VLM. These activity suggestions are often more expressive than the class categories that can be found in typical activity recognition datasets (e.g., Charades (Sigurdsson et al. 2018), Kinetics (Smaira et al. 2020)), and reflect the information already stored in the model, agnostic to the point of view. Our SM system for egocentric perception leverages these advantages, and also suggests future research directions in contextual reasoning that leverage existing language-based models without having to curate large annotated datasets.

3.3 Language-Based World-State History from Video

In order to provide language-based reasoning capabilities for open-ended question-answering, a key aspect of our system is to describe the observed states of the world in language, with the goal of creating a language-based world-state history (Fig. 4) that can be used as context to an LM. To this end, a component of our method generates Socratic image summaries of individual video frames (Sec. 3.3-A), that can then be concatenated (along with timestamps) to form an event log (illustrated at the top and middle of Fig. 4).

3.3-A. Socratic Egocentric Image Summaries. Given an image frame as input, this component generates a natural language summary (e.g., caption) of what is occurring in the image. Our system uses a Socratic approach with guided multimodal multi-model discussion to provide answers to 3 questions that describe the visual scene: "where am I?", "what do I see?", and "what am I doing?", which are then summarized into a single caption per image frame.

- **Where am I?** For place recognition, we use a VLM to



Places: clean room. Objects: shorts, jeans, shirt. Commonsense activities: getting dressed. Most likely: getting dressed.
SM (ours): I am getting dressed.

ClipCap: how to make a pair of jeans.

Places: kitchen. Objects: coffeemaker, waffle iron, kettle. Commonsense activities: making coffee, making waffles. Most likely: making coffee.
SM (ours): I am making coffee, waffles, and tea.

ClipCap: how to clean a stove with a brush.

Places: living room. Objects: remote control, television, netflix. Commonsense activities: watching netflix. Most likely: watching netflix.
SM (ours): I am watching netflix on the television.

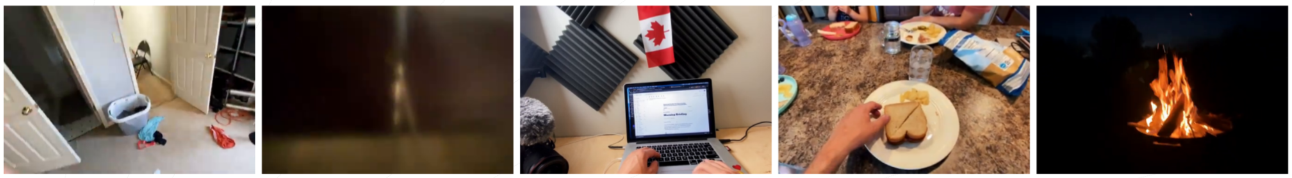
ClipCap: this is what the house looks like from the inside.

Places: kitchen. Objects: refrigerator, refrigerator, dishwasher. Commonsense activities: cooking, cleaning. Most likely: cooking.
SM (ours): I am cooking in a kitchen.

ClipCap: the refrigerator is full of food.

Places: kitchen. Objects: cooking spray, measuring cup, mixing bowl. Commonsense activities: measuring, mixing. Most likely: mixing.
SM (ours): I am mixing a recipe.

ClipCap: how to make a mason jar with a lid.



Places: clean room. Objects: closet, wardrobe, drawer. Commonsense activities: putting clothes away. Most likely: putting clothes away.
SM (ours): I am putting clothes away.

ClipCap: the dog's owner was left shocked when the cat jumped out of the way of the door.

Places: shower. Objects: light switch, curtain, mirror. Commonsense activities: turning on light, looking in mirror, showering. Most likely: showering.
SM (ours): I am showering and see the typical objects in a shower.

ClipCap: the video shows the man running away from the camera.

Places: home office. Objects: flag, poster, computer monitor. Commonsense activities: work on computer, look at flag, look at poster. Most likely: work on computer.
SM (ours): I am work on computer in home office.

ClipCap: the computer is now working on the screen.

Places: kitchen. Objects: sandwich, hamburger, kitchen & dining room table. Commonsense activities: eating, sitting. Most likely: eating.
SM (ours): I am eating a sandwich in a kitchen.

ClipCap: person, who is a student, said she was shocked when she saw the sandwich on the table.

Places: campsite. Objects: fireplace, torch, wood-burning stove. Commonsense activities: cooking, camping. Most likely: camping.
SM (ours): I am camping and can see a fireplace, torch, and wood-burning stove.

ClipCap: campfire in the night, slow motion.

Generated Language-Based World-State History from Egocentric Video

08:31 AM: Places: clean room. Objects: shorts, jeans, shirt. Activities: getting dressed. I was getting dressed.
10:17 AM: Places: kitchen. Objects: coffeemaker, waffle iron, kettle. Activities: making coffee. I was making coffee , waffles, and tea.
11:09 AM: Places: living room. Objects: remote control, television, netflix. Activities: watching netflix. I was watching netflix on the television.
01:17 PM: Places: staircase. Objects: stairs, hamster, human leg. Activities: ascending. I was ascending a staircase and see a hamster on the stairs and a human leg.
01:45 PM: Places: porch. Objects: package, porch, door. Activities: receiving. I was receiving a package.
03:24 PM: Places: kitchen. Objects: human hand, sink, human arm. Activities: washing dishes. I was washing dishes in a kitchen.
03:38 PM: Places: kitchen. Objects: refrigerator, refrigerator, dishwasher. Activities: cooking. I was cooking in a kitchen.
03:52 PM: Places: kitchen. Objects: cooking spray, measuring cup, mixing bowl. Activities: mixing. I was mixing a recipe.
04:24 PM: Places: shower. Objects: light switch, curtain, mirror. Activities: showering. I was showering and see the typical objects in a shower.
04:35 PM: Places: home office. Objects: flag, poster, computer monitor. Activities: work on computer. I was work on computer in home office.
05:16 PM: Places: kitchen. Objects: sandwich, hamburger, kitchen & dining room table. Activities: eating. I was eating a sandwich in a kitchen.
05:27 PM: Places: clean room. Objects: laptop, digital clock, desk. Activities: checking time. I was checking time and working on a laptop in a clean room.
06:31 PM: Places: field road. Objects: girl, worm, centipede. Activities: walking. I was walking in a field and see a girl and a worm.
07:01 PM: Places: yard. Objects: axe, axe, chainsaw. Activities: chopping wood. I was chopping wood.
07:12 PM: Places: campsite. Objects: fireplace, torch, wood-burning stove. Activities: camping. I was camping and can see a fireplace, torch, and wood-burning stove.

Open-Ended Reasoning

Recall Q&A

Q: Did I eat dinner today? **A**: Yes, I ate dinner today. **Explanation**: I was seen eating a sandwich in a kitchen at 5:27 PM.
Q: Did I drive today? **A**: No, I did not drive today. **Explanation**: I was at home all day.
Q: Did I do laundry today? **A**: No, I did not do laundry today. **Explanation**: I was putting clothes away in a closet and wardrobe and drawer in a clean room.

Contextual Reasoning Q&A

Q: What was I doing outdoors? **A**: I was chopping wood in a yard. **Explanation**: I need wood to keep my fireplace going.
Q: Why was I chopping wood? **A**: I was chopping wood because I needed to get a fire going in the fireplace. **Explanation**: I was camping and it was getting cold.
Q: Why did I go to the front porch today? **A**: I went to the front porch today to receive a package. **Explanation**: I saw on the porch a package and knew that I was expecting it.
Q: Why were my wife and I in the living room this evening? **A**: My wife and I were in the living room this evening because we were watching netflix. **Explanation**: we have a television in the living room and netflix is a service that we can watch on it.

Temporal Reasoning Q&A

Q: When did I last wash my hands? **A**: I last washed my hands at 3:38 PM. **Explanation**: I was washing dishes in a kitchen.
Q: How many times did I receive a package today? **A**: I received a package once today. **Explanation**: I was receiving a package at 03:24 PM.
Q: When did I last drink coffee? **A**: I last drank coffee at 10:17 AM. **Explanation**: I was making coffee in the kitchen.
Q: When did I last watch Netflix? **A**: I last watched Netflix at 7:32 PM on April 6, 2020. **Explanation**: I was watching Netflix in a living room.
Q: How many times did I use electronics today? **A**: I used electronics six times today. **Explanation**: I was working on a laptop in a clean room, checking time and watching netflix on a television, and streaming netflix on a computer monitor.
Q: When did I take a shower? **A**: I took a shower at 4:35 PM. **Explanation**: I was showering and see the typical objects in a shower.
Q: How long did I watch TV today? **A**: I watched TV for about 5 hours today. **Explanation**: I was watching netflix on the television for about 3 hours and then I watched a woman drink v in a living room for about 2 hours.

Subjective Q&A

Q: Was I happy today? **A**: I was generally happy today. **Explanation**: I was able to do some activities that I enjoy, such as watching netflix, working on my computer, and eating a sandwich.
Q: What was my favorite drink today? **A**: I drank wine in a living room with a woman. **Explanation**: I like to drink wine with friends.

Fig. 4: An instantiation of the SMs framework for open-ended reasoning with egocentric perception. SMs can generate meaningful structured captions (top) for egocentric images through Socratic dialogue between VLMs (green) and LMs (blue), and qualitatively perform well versus state-of-the-art captioning models such as ClipCap (Mokady, Hertz, and Bermano 2021). Key moments from egocentric video are summarized with SMs into a language-based world-state history (middle), which can be provided as context to an LM for open-ended question answering. Results (bottom) for generated answers (blue) and model explanations (blue) suggest SMs are fairly capable of performing a variety of reasoning tasks including answering binary yes or no questions, contextual and temporal reasoning questions, as well as subjective questions.

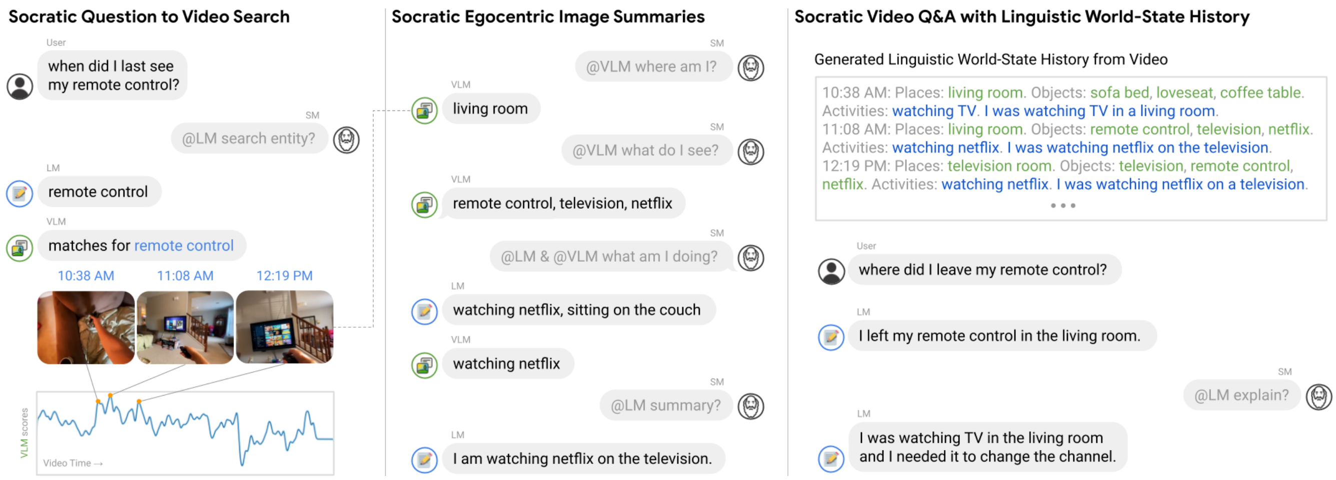


Fig. 5: Examples of guided multi-model exchanges (Socratic Models) for an egocentric perception system: (i, left) parsing a natural language question into search entities (with LM) to be used to find the most relevant key moments in the video (with VLM); (ii, middle) describing each key frame by detecting places and objects (VLM), suggesting commonsense activities (LM), pruning the most likely activity (VLM), then generating a natural language summary (LM) of the SM interaction; (iii, right) concatenating key frame summaries into a language-based world-state history that an LM can use as context to answer the original question.

rank Places365 (Zhou et al. 2016) scene categories against the image, with the top n candidates (out of 365) inserted into a prefix: "Places: {place1}, {place2}, {place3}.".

- **What do I see?** For object and people recognition, we use a VLM to rank OpenImages object categories (Kuznetsova et al. 2020) against the image, with the top m categories (out of 600) inserted into a second prefix: "Objects: {object1}, {object2}, {object3}."
- **What am I doing?** For activity recognition, we use a back-and-forth interaction between an LM and VLM: we first use an LM to infer the activities most related to the places and objects previously listed by the VLM (green):

```
Places: {place1}, {place2}, {place3}. Objects: {object1}, {object2}, {object3}. Activities: activity_a, activity_b, activity_c.
```

We find that generating candidate activities using an LM yields more suitable descriptions of egocentric activities and interactions with first-person video, than using standard activity recognition dataset categories (e.g., from Charades or Kinetics). Activity recognition datasets are often tailored to third person videos, and can only cover a partial subset of human activities, which instead can be more holistically captured through LM reasoning (Petroni et al. 2019) over the objects and places that the VLM perceives. For example, "receiving a package" is a common household activity not found in most datasets. After the LM generates candidate activities, these candidates are then fed back to the VLM and re-ranked to sort out the top k activities by relevance to the key image frame: "Activities: {activity1}, {activity2}, {activity3}."

This process of generating candidate activities from places and objects is one way of extracting commonsense from LMs as knowledge bases (Petroni et al. 2019). Continuing the Socratic dialogue further, this can be repeated likewise to generate new relevant objects (conditioned on activities and places), as well as new places (conditioned on objects and activities). One can iterate the procedure (LM generate, VLM re-rank, repeat) to populate the set of places, objects, and activities until equilibrium (i.e., no more new entities), which generally helps to cover a broader set of places and objects that expand beyond the initial seed categories from Places365 and OpenImages. For example:

```
If I am making making pancakes, objects that I am likely to see include: a frying pan, a spatula, a bowl, milk, eggs, flour, sugar, baking powder, butter, a plate, syrup.
```

Given the final set of places, objects, and activities, we use the LM to generate an overall first-person summary of what is happening in the image. Specifically, the prompt is:

```
I am in a place1, place2, place3. I see a object1, object2, object3. I am activity1. Question: What am I doing? Answer: I am most likely
```

The summarization process in general can capture more rich descriptions conditioned on the places, objects, and activities, and qualitatively seem to do well at ignoring irrelevant categories (i.e., denoising). For example:

```
I am in a nursing home, landfill, living room. I see a wine, wine glass, woman. I am drinking wine. Question: What am I doing? Answer: I am most likely enjoying a glass of wine with a friend or loved one.
```

However, while the LM's denoising capabilities can compensate for the shortcomings of the VLM, it is important to note that this may also cause unwanted ignoring of notable, but rare events (e.g., such as witnessing a purple unicorn, which may be ignored, but potentially it is Halloween).

Finding new ways in which such events can be indexed appropriately may be useful for downstream applications.

Egocentric Image Summary Results. On egocentric images, we show several qualitative examples of summaries generated by our system in Fig. 4, and compare them to results from a state-of-the-art image captioning model, ClipCap (Mokady, Hertz, and Bermano 2021). While state-of-the-art captioning models can perform reasonably over several of the images, we find that our system generally produces more relevant captions for a larger portion of the egocentric examples. Image captioning models are biased based on the datasets they are trained on, and have shown to perform poorly on egocentric images (Agarwal et al. 2020), which aligns with our observations. Relatively less research has been carried out specifically on egocentric image captioning (Fan, Zhang, and Crandall 2018). SMs can nevertheless produce reasonable captions without additional training on domain-specific data.

3.3-B. Adding Audio into Single-moment Summaries. In addition to using visual perceptual inputs, we may use a Socratic approach which engages perceptual inputs from audio as well, via an ALM (audio language model). Our example egocentric perception system uses Wav2CLIP (Wu et al. 2021a) as the ALM. Wav2CLIP is trained on 5-second audio clips from the VGGSound dataset (Chen et al. 2020), and is trained in a contrastive manner by aligning its audio encoder to the visual CLIP embeddings from video.

Incorporating an ALM like Wav2CLIP into our Socratic framework can provide an additional modality with which to perform zero-shot cross-modal reasoning, and this may help further improve inference beyond the vision-language-only case. Fig. 6 displays a driving example for which a visual-only summarization produced the less-than-desirable summary: "I am climbing a staircase, and I may see a hamster or human leg" with the incorrect propogation of the false detection of a hamster and human leg.

To perform audio-aided single-moment summarization, we first run image-based summarization as described previously, but we then prompt the LM to suggest sounds that it may hear, given the visual context, via "⟨visual single-image summary⟩. 5 Possible Sounds:". For the example in Fig. 6 an example prompt, which has already gone through multiple rounds of Socratic dialogue to be generated, together with completion by the LM is:

```
Places: staircase. Objects: stairs, animal, mammal,
hamster, human leg. Activities: climbing. 5 Possible
Sounds: footsteps, creaking stairs, someone calling
your name, a dog barking, a centipede crawling.
```

These auditory entities expressed in language can then be ranked by the ALM. In this moment of the video, the sound of footsteps can be faintly heard in the background, and in this case the ALM provides a correct detection of ranking footsteps as the most likely sound. This ranking can then be incorporated into a prompt for the LM to provide the single-image summary, for example:

Fig. 6: Example frame and corresponding (centered) 5-second audio clip which provide the driving example for Sec. 3.3-B, i.e., adding in ALMs into Socratic dialogue to improve single-moment summarization. Note that this waveform mostly represents the background piano music, but the system is still able to rank correctly that footsteps as the highest sounds relative to others in the LM-suggested candidate set.

```
I am in a: {place}. I see a: {object1}, {object2},
{object3}, {object4}, {object5}. I think I hear
{sound1} I am: {activity}. Summary: I am most likely
```

As above, incorporating "I think I hear footsteps" into the summary and prompting this to the LM provides the completion: "**climbing a staircase, and I may hear footsteps.**" In this case, this summary result is preferable to the mentioned single-image caption without sound.

While this example demonstrates in a certain case the utility of audio-informed summaries, overall in egocentric video, with a variety of background noise, we find that Wav2CLIP can provide reasonable detections for certain language-represented auditory entities such as 'baby babbling' and entities to do with running water, but do not provide as robust detections as CLIP. Also, while there are many advantages to the specific Wav2CLIP approach, including its use of the CLIP embedding space, a major downside is that the training process is "blind" to hearing things that cannot be seen. Accordingly, for the rest of demonstrations shown, we simply build world-state history from VLM-LM interactions alone. We expect however that with further attention to model approaches, and scaling of audio-language datasets, approaches like Wav2CLIP will increase in robustness. We also show an additional application (Sec. 3.4) of audio, for audio retrieval. In that case, only a single auditory search entity is required in order to enable a useful application, and so it can be easier to verify that it is a sufficiently robustly-detected entity.

3.3-C. Compiling a Language-Based World-State History

Our system compiles the image summaries from each key video frame into a language-based world-state history. Since the total number of frames in the video may be large, compiling a summary for every individual frame would create text that is too large (too many tokens) to be processed directly by an LM as context for Q&A. Accordingly in this work, we propose solutions that sparsify and/or condense language-based world-state histories (e.g., via search-based methods) into practically usable context sizes for reasoning. In particular, we explore two methods of identifying "key moments" in videos for summarization: (i) uniform sampling over time, and (ii) video search (image or audio retrieval) for on-the-fly compilation of context.

The first method, uniform sampling, is straightforward and compiles a world-state history from Socratic summaries of video frames sampled at fixed time intervals. This can also be condensed hierarchically using recursive linguistic summarization (Wu et al. 2021b), to fit even dense sampling into usable LM-context sizes. However, while broadly indiscriminate, uniform sampling may not have sufficient temporal resolution to capture important spontaneous events in the video (such as adding salt to the pot while cooking soup in the kitchen).

Hence the second method, identifying key moments with video search, uses a VLM or ALM to search for entities most relevant to the question, which can more precisely index the frames in which the subject appears. Specifically, our instantiation of SMs for this component parses a natural language question with an LM into several search entities to be used to find key frames in the video. For example, the question "did I drink coffee today?" yields a search entity "**drink coffee**" that is then used with language-conditioned video search to index the most relevant n key frames of "drink coffee" in the video. The LM categorizes the search, which can be image-based (VLMs) or audio-based (ALMs), e.g., for language-conditioned auditory recall questions ((Oncescu et al. 2021)) like "why was my wife laughing today?" . While search-based indexing of key moments can be useful for finding spontaneous events, this method for generating context can also provide disadvantages for downstream Q&A if the answer to the question depends on events that are not directly related to the search subject. For example, "why was I chopping wood today?" returns key frames related to "chopping wood", but does not return the key frames after the event related to making a campfire. On the other hand, if uniform sampling is employed and the campfire events are captured by the summary, then the LM can successfully return the answer "**I was making a campfire.**" Choosing which method to use for compiling the language-based world-state history may depend on the application.

Language-based World-state History Results. Fig. 4, middle, shows results generated by our system. The specific event log shown in Fig. 4 has been trimmed down for space considerations, but is representative of the type of event logs that may be generated without manual curation. These event logs are used as context to enable LM open-ended reasoning on video, as demonstrated in the next section.

3.4 Open-Ended Reasoning on Egocentric Video

In this section we describe a few examples of how the Socratic Models framework can be used to perform open-ended multimodal-informed completion of text prompts, conditioned on egocentric video (examples in Fig. 2). There are of course limitations to they can provide, but our demonstrated examples suggest that we can already today often generate compelling answers to open-ended reasoning tasks, at a scope that is beyond what we are aware is possible today with available methods. Of course, the answers may also inherit undesirable characteristics from the component models, such as an LM that is overconfident even when wrong. It is our hope that our results may help inspire work on preparing even more comprehensive video understanding datasets for the community, to assist further assessment.

Our example system uses a language-based world-state history generated through Socratic multi-model discussion (Sec. 3.3), and provides this as context to an LM to enable open-ended reasoning on egocentric videos. Open-ended text prompts from a user, conditioned on an egocentric video, can yields three types of responses: a text-based response, a visual result, and/or an audio clip. These latter two provide examples that open up the capabilities of the system to respond not only with text-based responses, but also respond with video snippets themselves, which may be a higher-bandwidth way to respond to user requests (*"a picture is worth a thousand words"*). The specific composition of our system is of course just one example – overall, the modularity of the Socratic approach makes it easy to compose together foundation models, zero-shot, in a variety of ways to provide a spectrum of multimodal reasoning capabilities.

The demonstrated tasks include (i) summarization, (ii) open-ended Q&A, (iii) forecasting, (iv) corrections, and (v) video search for either visual or audio cues. These tasks have predominantly been studied in isolation in the research community – but our example results with SMs suggest they can be subsumed under the same unified language-based system for multimodal reasoning. ***Descriptions and results*** for each of (i)-(v) are shown below.

(i) Summarization can be implemented by prompting an LM to complete the excerpt "{world-state history} Summary of my day:" to which it can respond with outputs like "**I slept in a bed, made coffee, watched TV, did laundry, received a package, bench pressed, showered, ate a sandwich, worked on a computer, and drank wine.**" Since the language-based world-state history is constructed with summaries of visual content, it carries contextual information that can be complementary to what is found in closed captions (e.g., speech and dialogue, explored in Sec. 5). Summarizing egocentric videos enables a number of applications, including augmenting human memory to recall events, or life-logging of daily activities for caregiver assistance. Our system draws similarity to early work in the area involving text-based summarization and identifying key frames (see (Barbieri, Agnihotri, and Dimitrova 2003) for an early survey and (Del Molino et al. 2016; Apostolidis et al. 2021) for more recent surveys).

(ii) Open-ended Q&A can be implemented by prompting the LM to complete the template: "{world-state history} Q: {question} A:". We find that LMs such as GPT-3 can generate surprisingly meaningful results to binary yes or no questions, contextual reasoning questions, as well as temporal reasoning questions. As in (Yang et al. 2021) we can further

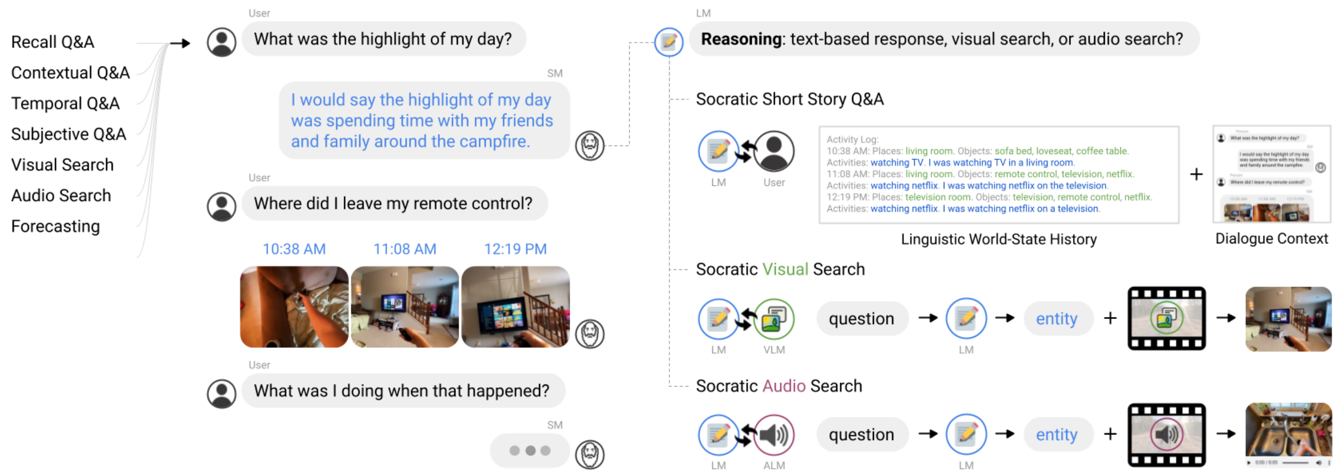


Fig. 7: SMs can interface with the user through dialogue and perform a variety of tasks (formulated as Q&A) with egocentric video: sorting reasoning questions by their output modalities e.g., text-base responses, images from visual search, video snippets from audio search. Depending on the modality, each question can pass through a different sequence of Socratic interactions between the LM, VLM, and ALM.

prompt the LM to *explain the answer* by adding "This is because:". We find that the accuracy of the answers and explanations remain largely conditioned on whether the necessary information can be found within the world-state history. This suggests that the quality of the language-based reconstructions of the videos (e.g., via key frame sampling and captioning in this work) is central to the approach.

We show several qualitative examples of free-form question answering using our SM system on egocentric video in Fig. 4, bottom, Fig. 5, and Fig. 7 generated using a first-person POV video[2] as input.

Recall Questions. SMs can perform simple retrieval of events. For example, "did I eat dinner today?", yields a response "**yes I ate dinner today.**" along with an explanation "**I was seen eating a sandwich in a kitchen at 5:27 PM.**" which points to the key frame that was captioned with the sandwich in hand. Another example that involves contextual reasoning to recall events is "what was I doing outdoors?" to which the system responds "**I was chopping wood in a yard.**" Likewise, if the entities described in the question do not appear in the world-state history, such as "did I drive today?" the system can respond with a negative answer: "**no, I did not drive today.**" with an explanation "**I was at home all day.**" This capability expands beyond standard video search, which might only return nearest neighbor video frames, without a natural language response (or a negative response).

The performance of recalling events largely depends on the relevance of the language-based world-state history to the question. We find that recall-type questions work best with world-state history logs that are compiled by using search-based key frame indexing (see Sec. 3.3-B). The system can still return negative responses, since the captioning of the key frames are not influenced by the question.

Temporal Reasoning. SMs can answer questions related to time by appending timestamps to each key moment in the world-state history. By associating image summaries to times of the day, this allows answering questions that time index various activities. For example "when did I last drink coffee?" can return the last time drinking coffee was mentioned in the log, with a full response "**I last drank coffee at 10:17 AM**" and an explanation "**I was making coffee in the kitchen.**" The system can also count events, for example when asked "how many times did I receive a package today?", the system will respond appropriately "**I received a package once today.**" with an explanation "**I was receiving a package at 3:24 PM**". We find that a common failure mode for these types of questions is that the system tends to over-count, especially as a reaction to false positive VLM detection results that get surfaced into the world-state history. For example, asking "who did I interact with?" would yield "**woman, hamster**" where hamster was a false positive prediction from CLIP. These issues become more prominent with search-based key frame sampling, as a byproduct of an inability to distinguish neighboring local argmaxes of the same event from each other.

Cause and Effect Reasoning. SMs can answer questions about cause and effect relationships between events, conditioned on that all the events appear in the world-state history. For example, when asked "why did I go to the front porch today?" the system would respond "**I went to the front porch today to receive a package.**" and an explanation "**I saw on the porch a package and knew that I was expecting it.**" These types of questions are exciting because they suggest opportunities for prompting logical deduction of events. However, since information about both the cause and the effect needs to be in the world-state history, the quality of results remains highly dependent on the key frame sampling strategy used to compile it (Sec. 3.3-B). Uniform gives an

[2]Examples on https://youtu.be/-UXKmqBPk1w used with permission from Cody Wanner.

unbiased account of events, and is currently the best variant for this form of reasoning. More targeted construction of the world-state history with search based key frames can sometimes miss frames that capture the answer to the question.

Subjective Reasoning. SMs can also answer more subjective questions, such as "was I happy today?" or "what was my favorite drink today?". Without additional context, these questions rely on biases from the LM's dataset – which could have negative consequences, and should be managed carefully with additional mechanisms for safety and groundedness (Thoppilan et al. 2022). The full personalization of these subjective questions are likely to be conditioned on whether a better context can be constructed of prior user behaviors related to the question.

(iii) Forecasting of future events can be formulated as language-based world-state completion. Our system prompts the LM to complete the rest of an input event log. Timestamps of the predictions can be preemptively specified depending on the application needs. The completion results are generative, and are more broad than binary event classification (e.g., (Lei et al. 2020)). Example completion results (also shown in Fig. 2):

```
1:46 PM: I am eating a sandwich in a kitchen.
2:18 PM: I am checking time and working on a laptop
in a clean room.
2:49 PM: I am buying produce from a grocery store or
market.
3:21 PM: I am driving a car.
4:03 PM: I am in a park and see a playground.
4:35 PM: I am in a home and see a television.
```

Few-shot prompting the LM with additional examples of prior event logs most similar to the current one is likely to improve the accuracy of the completion results. Without additional context, these results are again biased towards typical schedules seen by the LM across Internet-scale data.

To a certain extent, this forecasting capability extends and generalizes the traditional topic of activity forecasting in computer vision. In the research community, activity forecasting has been often formulated as an extension of action classification, tracking, or feature generation: Given a sequence of image frames, they directly predict a few categorized actions (Ryoo 2011; Hoai and De la Torre 2014; Rhinehart and Kitani 2017), human locations (Kitani et al. 2012), or image features (Vondrick, Pirsiavash, and Torralba 2016) to be observed in the future frames. In contrast, Socratic Models with LMs enables generating more semantically interpretable descriptions of future events, conditioned on multimodal information.

(iv) Corrections. SMs can be prompted to incorporate human feedback in the loop as well, which could be useful for interactive language-based systems. For example, given image captions generated from an VLM and LM:

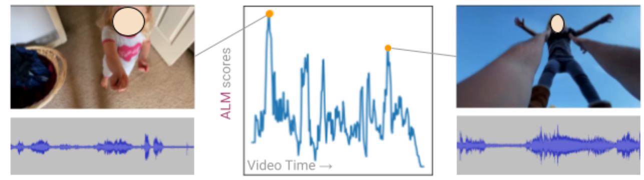


Fig. 8: Example zero-shot language-prompted auditory retrieval (shown: top 2 results) in response to "what did my daughter's laugh sound like today?", for which an LM identifies the audio search query of "daughter's laugh", and an ALM (Wav2CLIP) is used for audio retrieval. The top (left) retrieval is only partially correct, returning a video clip involving the daughter but not laughter. The second (right) retrieval is correct, from a moment of playing (getting tossed into the air). Faces obscured for privacy.

```
Context: Where am I? outdoor cabin, campsite, outdoor
inn. What do I see? fire, marshmallow, fire iron,
hearth, fireside, camp chair. What am I doing?
Commonsense suggests: roasting marshmallows, sitting
around the fire, chatting. Most likely: sitting
around the fire.
Original Summary: I am camping and enjoying the
company of my friends around the fire.
Corrections: It was actually my family, not friends,
sitting around the fire.
Corrected Summary: I am camping with my family and
enjoying the company of them around the fire.
```

(v) Video Search: Image or Audio Retrieval. Our SM system can also return additional modalities (images, audio) as answers to questions, by simply few-shot prompting the LM to classify a target modality based on the input question. For example, "where did I leave my remote control" can map to image search using VLM features for "remote control" while "what did my daughter's laugh sound like today?" can map to natural-langauge-queried audio search ((Oncescu et al. 2021)) using ALM features for "daughter's laugh" (Fig. 8). This can be useful for some applications (e.g., AR) in which the user may find the retrieved modality to be more useful than a natural language response. Our approach for this uses an LM to parse a search entity from the question to index key video frames. This is done with several few-shot examples provided as context. For example, the question "when did I last wash my hands?" yields a search entity "**wash my hands**" that is then used with video search to index the most relevant n key frames of "wash my hands" in the video. Specifically, our system runs video search by ranking matching CLIP or Wav2CLIP features of the entity text against all video frames, and returning the top n local maximums. For each frame, the features can either be image features or audio features (e.g., from the surrounding 5 seconds with Wav2CLIP) – where the LM few-shot categorizes which domain to use for any given question. This can be thought of as calling different subprograms for hierarchical search.

Limitations. Overall, our results suggest that SMs are capable of generating meaningful outputs for various egocentric perception tasks via visual contextual reasoning – but its limitations also suggest areas for future work. For example, a primary bottleneck in the Q&A system is that it relies on the richness (i.e., recall) and quality (i.e., precision) of the event log. This likely could be improved with better image and audio detectors or captioning systems (Gu et al. 2021). Also, we find that the used Wav2CLIP may provide satisfactory results for certain categories in audio retrieval, but we currently do not involve it in generating the event log, since its robustness and range of open-language detection is not at the same level as CLIP. This seems addressable with further approaches and scaling of datasets in the audio-language domain.

Additionally, accurate response to cause and effect reasoning questions also require relevant key moments to be reflected in the event log – which points to open ended questions on how to achieve better key frame sampling (beyond the simple baselines that we have demonstrated). Finally, the dialogue between the different models are fairly structured with manually engineered prompts. It may be interesting to investigate more autonomous means of achieving language-based closed loop discussions between the models until a commonsense consensus is reached.

3.5 Scaling Up Socratic Video Search

The search algorithms of the SMs, which may be used both for compiling world-state history (Sec. 3.3-C) and for video search retrieval (Sec. 3.4) rely on the matching procedure conducted in the corresponding latent space (e.g. VLM features of the text snippet against these of the video frames). This can be abstracted as dot-product-maximization key search in the given key-dataset. In practice, if the key-dataset is large (e.g. long videos) a naive linear search is prohibitively expensive. We propose several solutions to this problem.

MIP-Search. The first observation is that several data preprocessing techniques applied in the so-called *maximum inner product* (MIP) search can be directly used to reorganize the keys (e.g. latent representations of video frames) to provide sub-linear querying mechanism for the incoming text snippet (see: (Abuzaid et al. 2019)). Those include pruning and various indexing techniques, such as LSH-hashing (Shrivastava and Li 2014). In the hashing approach, a collection of hash-tables, indexed by the binarized representations of the hashes is stored with different entries of the hash table corresponding to the subsets of keys producing a particular hash. There are several cheap ways of computing such hashes, e.g. *signed random projection* (those in principle linearize the angular distance, but every MIP task can be translated to the minimum angular distance search problem). The querying is then conducted by searching for the most similar hash-entries in the hash-tables and then performing linear search only on the subsets of keys corresponding to these entries to obtain final ranking.

Associative Memories. The above approach provides sub-linear querying mechanism, but does not address the space complexity problem. In the scenario of strict memory requirements, we propose to leverage recently introduced techniques on linear attention (Choromanski et al. 2021b) combined with *modern continuous associative memory* (MCAM) models (Ramsauer et al. 2021). MCAM models are de facto differentiable dictionaries (with provable few-shot retrieval) that can be thought of as energy-based models using negated exponentiated latent-representations-dot-product energy for the *exponential* storage capacity. A naive computation of such an energy still requires explicitly keeping all the patterns (which is exactly what we want to avoid), but this can be bypassed by applying the linearization of that energy (which effectively is just the negated sum of the softmax kernel values) with the FAVOR+ mechanism used in linear-attention Transformers, called *Performers* (Choromanski et al. 2021b). This modification has several advantages: (1) it makes the size of the dictionary completely independent from the number of the implicitly stored patterns; the size now scales linearly with the number of random features used for energy linearization, (2) it provides a *constant-time* querying mechanism at the price of compressing all the patterns (and thus losing some information).

Random Feature Trees. The other approach, that combined the ideas from both MIP-search and linear attention systems, leverages the so-called *random feature tree* (RFT) data structure (Rawat et al. 2019). This approach relaxes the MIP-search to sampling from the linearized softmax distribution via FAVOR+ (Choromanski et al. 2021a). Sampling from such a linearized distribution can be done in time logarithmic in the number of samples via RFT which is a balanced tree with leaves corresponding to latent representations of video frames and nodes encoding representations of the subsets of keys (e.g. the video frames) defined as sums of the random feature transforms of the keys.

4 Socratic Internet Data Image Captioning

The SMs framework can also be used to generate text captions for generic Internet images with a guided multi-model exchange between a VLM and LM. We describe an example system in Sec. 4.1 and demonstrate results in Sec. 4.2.

4.1 System: Image Captioning on Internet Data

Overall, our example SMs system for Internet image captioning is extremely similar to how we perform single-image captioning in our egocentric system, but (i) adapted for Internet images rather than ego-centric images, and (ii) adapted such that the “final task” is the generation of a single image caption, rather than open-ended tasks based on text-prompted completion.

First, similar to the process of generating egocentric image captions, we may prompt the VLM to zero-shot detect visual entities across different categories of language. As with the egocentric system, we return top matching place categories and object categories. For Internet data, we use Tencent ML-Images (Wu et al. 2019) for object entities. We also choose to detect the image type from the set {*photo,*

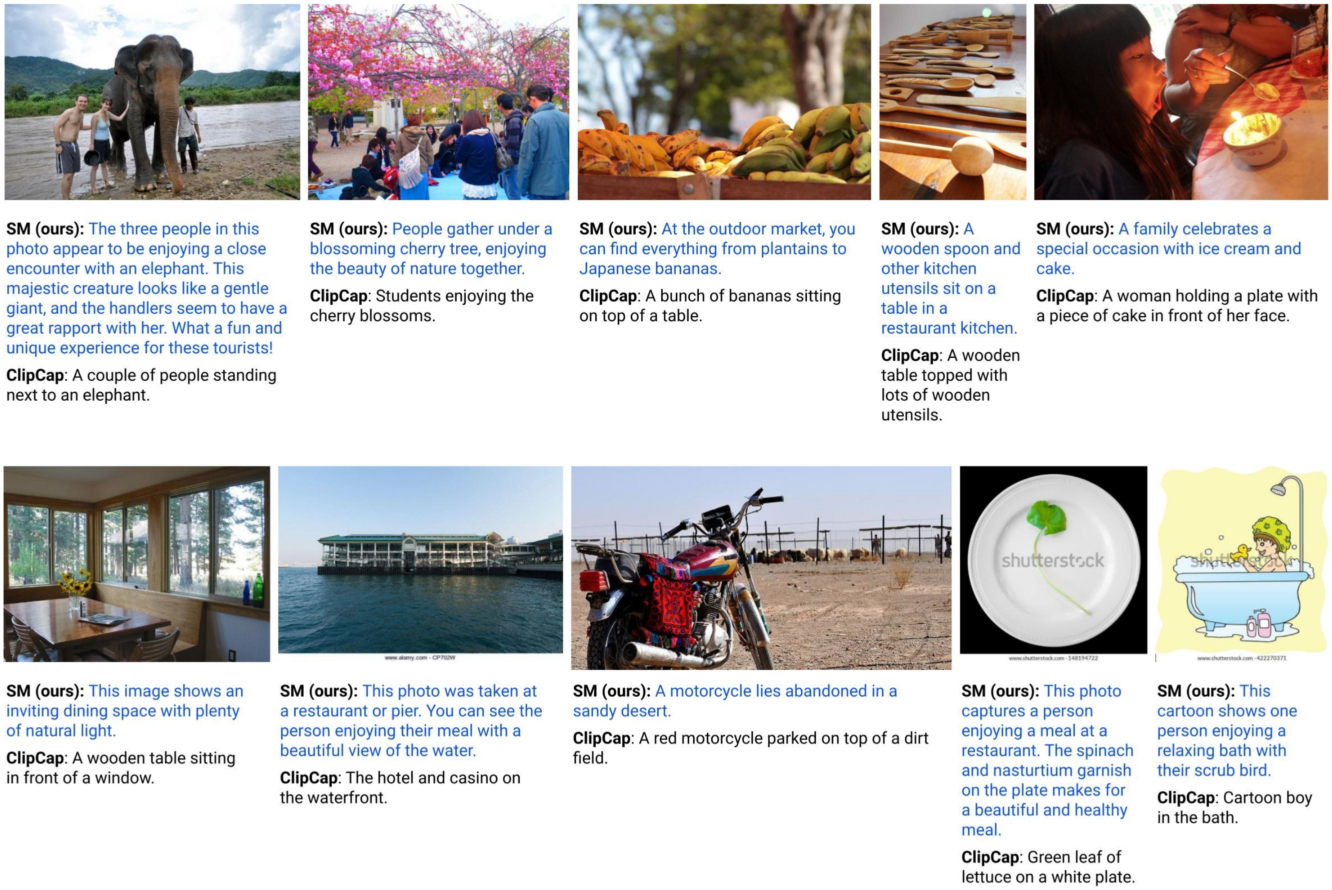


Fig. 9: The SMs framework with a VLM and LM can be used to *zero-shot* generate captions for generic Internet images, and can be as expressive as *finetuned-for-captioning* state-of-the-art models e.g., ClipCap (Mokady, Hertz, and Bermano 2021).

cartoon, sketch, painting} and the amount of people from the set {*are no people, is one person, are two people, are three people, are several people, are many people*}. For generic Internet images, which are not necessarily real photos, and very often are taken by people of people, we find that these additional contextual pieces of information help generate better captions. The various VLM detections give: "Places: {place1}, {place2}, {place3}. Objects: {object1}, {object2}, {object3}. Image type: {image_type}. People result: {people_result}."

Next, given the VLM detections of various visual entities, we can then prompt the LM to generate several (n) candidate captions. For this step, we employ a non-zero sampling temperature (we find 0.9 gives good results) in order to give sufficient variety of results across the n options, but still generate reasonably likely options.

```
I am an intelligent image captioning bot. This image
is a {img_type}. There {num_people}. I think this
photo was taken at a {place1}, {place2}, or {place3}.
I think there might be a {object1}, {object2},
{object3},... in this {img_type}. A creative short
caption I can generate to describe this image is:
```

Examples of an actual prompt and generated sample captions, for example for the bottom-left photo of Fig. 9 is:

```
I am an intelligent image captioning bot. This image
is a photo. There are no people. I think this photo
was taken at a indoor bow window, dining room, or
interior balcony. I think there might be a double-hung
window, casement window, sliding window, pivoting
window, breakfast area, breakfast nook, dining area,
storm window, storm sash, dining room, bay window,
bow window, lancet window in this photo. A creative
short caption I can generate to describe this image
is: ... (run n times, examples shown below)

• Looking through the window to a beautiful view.
• This serene and stately room is centered around
  a beautiful window, which floods the space with
  natural light. The double-hung window allows for
  ventilation, while the storm sash protects against
  inclement weather. The perfect spot for a lazy
  Sunday brunch or an intimate dinner party.
• ...
```

Lastly, after the LLM has generated several captions, they are subsequently ranked by the VLM, and the highest scoring caption is returned – in the example above, this is: "This image shows an inviting dining space with plenty of natural light."

4.2 Results: Image Captioning on Internet Data

Fig. 9 shows several qualitative comparisons with ClipCap (Mokady, Hertz, and Bermano 2021), a state-of-the-art method for image captioning specifically trained via finetuning on image captioning. We show zero-shot generation of captions on the set of images displayed in ClipCap's code release.[3] We make a colab available to generate these results.[4]

Overall, our results (Fig. 9) show that the Socratic Model framework can be adopted to provide often convincing results for image captioning via the creativity of the LM combined with the visual grounding of the VLM. While these results are promising, the degree to which visual details are provided in the captions is largely limited by the capabilities of the VLM. For example, attributes (e.g., color of a shirt, a person's facial expression, or the spatial relationships between objects) are details not often captured in our particular system, which relies more on the contextual image classification capabilities of the VLM. Future work may explore open-vocabulary object detectors (Gu et al. 2021) as a means to recover salient details for more rich captions. It is also important to note that the system may generate captions that reflect unwanted biases found in the Internet-scale data that it is trained on, and should be used with caution (and checked for correctness) in downstream applications.

5 Socratic Video-to-Text Retrieval

We also adapt the Socratic Models framework to the task of video-to-text retrieval, a common video understanding task with clear and quantitative metrics. We describe an example SM-based system (Sec. 5.1) which uses a guided multi-model exchange between a VLM, a causal LM, a speech-to-text (ALM) model, and a masked LM, that achieves state-of-the-art performance for zero-shot video-to-text (i.e., caption) retrieval on the popular MSR-VTT dataset (Sec. 5.2).

5.1 System Overview: Socratic Video-to-Text Retrieval

Our approach uses SMs to augment Portillo-Quintero et al. (2021) with commonsense information from the audio and language domains. Portillo-Quintero et al. (2021) computes a similarity measure between the average VLM (i.e., CLIP) features of the image frames from a video, and the CLIP text features of a caption. This zero-shot method can be used directly for MSR-VTT video-to-text retrieval via one-to-many nearest neighbor matching.

We can improve upon their method through a combination of speech-to-text ALMs together with LM-based commonsense reasoning. First, we transcribe the audio from all

[3]https://github.com/rmokady/CLIP_prefix_caption

[4]Note that due to the non-zero temperature used for sampling from the generative language model, results from this approach are stochastic, but comparable results are producible.

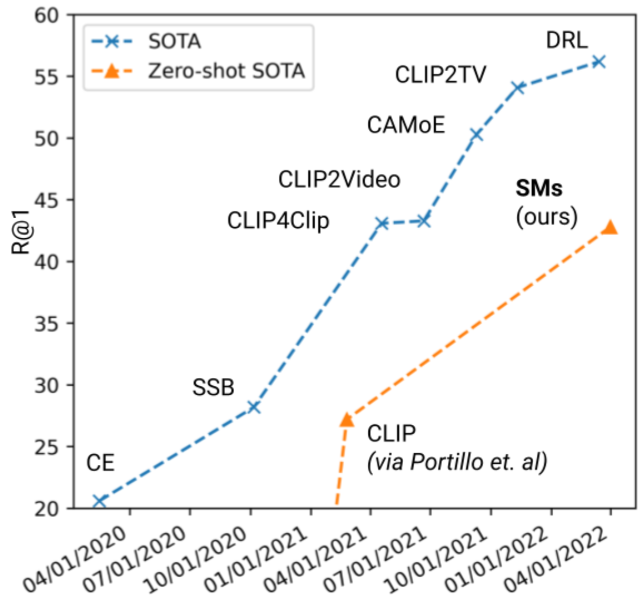


Fig. 10: Evolution over time of state-of-the-art (SOTA) results for video-to-text retrieval, with R@1 metric, on the popular MSR-VTT (Xu et al. 2016) 1k-A (Yu, Kim, and Kim 2018) dataset. See Tab. 1 for additional information on these methods.

videos with speech-to-text ALMs (Bapna et al. 2022) (also called ASR, or automatic speech recognition), using the publicly-available Google Cloud speech-to-text API.[5] Although raw transcripts may be challenging to incorporate into meaningful improvements for video/caption retrieval, we may leverage reasoning capabilities from large LMs in order to usefully harness the transcripts. For videos with sufficiently long transcripts, we summarize the content with an LM (e.g., GPT-3) using the following prompt:

```
I am an intelligent video captioning bot.' I hear
a person saying: "{transcript}". Q: What's a short
video caption for this video? A: In this video,"
```

We then compute the similarity scores of the generated summary to the set of captions with a masked LM (e.g., similarity between sentence embeddings from RoBERTa (Liu et al. 2019b)), and use those scores to re-weight the CLIP-based ranking from Portillo-Quintero et al. Specifically, for videos with sufficiently-long transcripts (we use $\geq$100-character transcripts), the matching score is: $\big(\textit{CLIP}\,(\text{caption}) \cdot \textit{CLIP}\,(\text{video}')\big) \times \big(\textit{RoBERTa}\,(\text{caption}) \cdot \textit{RoBERTa}\,(\textit{GPT-3 with prompt}\,(\textit{Speech2Text}\,(\text{audio}')))\big)$, where $\cdot$ represents normalized dot product of embeddings, and $\times$ represents scalar multiplication. For a given video, if there is no audio or the transcript is too short, we resort to exactly the method of Portillo-Quintero et al., which computes matching scores with $\textit{CLIP}(\text{caption}) \cdot \textit{CLIP}(\text{video}')$.

[5]https://cloud.google.com/speech-to-text. Key used parameters include 'model=video' and 'use_enhanced=True'. At 0.006 cents per 15 seconds, this represents an estimated speech-to-text processing cost of under 25 cents (USD) for all MSR-VTT test data.

		MSR-VTT 1k-A				MSR-VTT Full					
Category	Method	R@1↑	R@5↑	R@10↑	MdR↓	R@1↑	R@5↑	R@10↑	MdR↓	Audio	CLIP enc.
	JEMC (Mithun et al. 2018)					12.5	32.1	42.4	16.0	yes	
	Collaborative Experts (Liu et al. 2019a)	20.6	50.3	64.0	5.3	15.6	40.9	55.2	8.3	yes	
	SSB (Patrick et al. 2020)	28.5	58.6	71.6	3.0					no	
	CLIP4Clip (Luo et al. 2021)	43.1	70.5	81.2	2.0					no	ViT-V/32
Finetuning	CLIP2Video (Fang et al. 2021)	43.5	72.3	82.1	2.0	**54.6**	**82.1**	**90.8**	**1.0**	no	ViT-V/32
	DRL (Wang et al. 2022), ViT-B/32	45.3	73.9	83.3	2.0					no	ViT-V/32
	CAMoE (Cheng et al. 2021)	49.1	74.3	84.3	2.0					no	ViT-B/32
	CLIP2TV (Gao et al. 2021)	54.1	77.4	85.7	**1.0**					no	ViT-B/16
	DRL (Wang et al. 2022), ViT-B/16 + QB-n	**56.2**	**79.9**	**87.4**	**1.0**					no	ViT-B/16
	SSB (Patrick et al. 2020), zero-shot	8.7	23.0	31.1	31.0					no	
Zero-shot	CLIP via *Portillo-Quintero* (2021)	27.2	51.7	62.6	5.0	40.3	69.7	79.2	**2.0**	no	ViT-B/32
	SMs (ours)	**42.8**	**62.6**	**70.6**	**2.0**	**44.7**	**71.2**	**80.0**	**2.0**	yes	ViT-B/32

Tab. 1: Video-to-text retrieval results on MSR-VTT (Xu et al. 2016) dataset, both on the popular 1k-A (Yu, Kim, and Kim 2018) subset and the original 'full' test set. Differentiated are methods which train on the MSR-VTT dataset (*finetuning*), compared with *zero-shot* methods, which do not. Also noted: whether the methods use audio channels, and if CLIP (Radford et al. 2021) is used, which CLIP encoder is used. See Fig. 10 for the chronology of the SOTA across each category.

Here, the Socratic interaction lies mainly between the ALM (speech-to-text) to the commonsense LM (GPT-3 to summarize the transcriptions), and between the commonsense LM to the ranking based system that is a combination of the VLM (CLIP) and the masked LM (RoBERTa). Note that we may also prompt LMs (in this case, via multiple-choice) to determine if one caption is a better fit than another for a given video. However, for this specific task and dataset, with thousands of possible answers to choose from, the numerical ranking provided by embedding similarity scores provides a practical solution rather than relying on thousand-way multiple-choice commonsense reasoning.

5.2 Results: Socratic Video-to-Text Retrieval

	Long-transcript subset of… MSR-VTT 1k-A			
Method	R@1↑	R@5↑	R@10↑	MdR↑
CLIP via *Portillo-Quintero* (2021)	28.2	49.9	60.3	6.0
SMs (ours)	**55.0**	**71.6**	**76.3**	**1.0**
	MSR-VTT Full			
CLIP via *Portillo-Quintero* (2021)	41.5	69.6	77.4	2.0
SMs (ours)	**54.9**	**74.0**	**79.9**	**1.0**

Tab. 2: Evaluation on MSR-VTT for video-to-text retrieval with *long-transcript videos*: the subset of videos for which our SMs method used transcripts on 1k-A (n=451 out of 1,000) and full (n=1,007 out of 2,990). On these subsets, we evaluate Portillo-Quintero et al. (2021) vs. our method. Outside this subset, we resort to Portillo-Quintero et al.

For video-to-text evaluation we use the MSR-VTT dataset (Xu et al. 2016), which as noted in other recent works (Gao et al. 2021; Cheng et al. 2021) is the most popular benchmark dataset for the task of video-to-text retrieval. Like other recent works (Gao et al. 2021), we focus our results on this dataset. One of the reasons this is a good task and dataset for generally testing the value of the SMs approach is that there is already a strong zero-shot baseline, provided by Portillo-Quintero et al. (2021), which uses CLIP by itself, but does not use the Socratic method: there is no multi-model exchange, and no LMs are used. Additionally, this task provides a great opportunity to incorporate another type of modality – speech-to-text from audio data. We compare our method both with zero-shot methods, and with finetuned methods specifically trained on MSR-VTT.

Results show that our method sets a new zero-shot state-of-the-art for video-to-text retrieval on MSR-VTT (Tab.1), both on the "1k-A" and "full" test sets. Since our demonstrated system uses exactly the method of Portillo-Quintero et al. (2021) for its processing of CLIP features but additionally incorporates LLM reasoning on speech-to-text transcripts, the increased measured performance of our method (i.e. 27.2 → 42.8 R@1 on MSR-VTT 1k-A) directly reflects the additional benefit of incorporating language-based multimodal reasoning. Additionally, to keep the comparison between our method and Portillo-Quintero et al. (2021) as direct as possible, we maintain the usage of their precomputed CLIP features[6] from ViT-B/32. Given results from other recent methods (Tab. 1) it seems likely we may be able to improve our performance by switching to ViT-B/16, or other recent more-performant VLM models (Zhai et al. 2021).

As shown in Table 2, if we look at only the *long-transcript videos*, i.e. the videos for which our method used a transcript, then we especially see an increase in performance – on MSR-VTT 1k-A, R@1 almost doubles, from 28.2 to 55.0, for our method compared to Portillo-Quintero et al. (2021). Further, although it is on only a subset of the test set, note that this R@1 metric achieved of 55.0 is roughly comparable to the R@1 of the best *finetuned*-SOTA method, DRL (Wang et al. 2022) on the entire 1k-A dataset, with 56.2 R@1 (Tab. 1). If we assume that, for visual-only methods, the videos with-or-without transcripts are of roughly equal difficulty from a visual-only retrieval perspective, this suggests that on internet videos with sufficient spoken language

[6]https://github.com/Deferf/CLIP_Video_Representation

present, our method for zero-shot video-to-text retrieval can nearly match the *finetuned*-SOTA method for video-to-text retrieval.

Note that instead of *video-to-text* retrieval, but rather on *text-to-video* retrieval, a recent method (Li et al. 2022) has shown strong zero-shot results. Other methods have also attempted zero-shot on MSR-VTT *text-to-video* retrieval (Xu et al. 2021; Miech et al. 2020; Bain et al. 2021), but these have all been outperformed by Portillo-Quintero et al. (2021). Our method may be adapted as well to *text-to-video*, but due to our use of transcripts on only a subset of the videos, unlike in video-to-text, this creates an asymmetry which may require an unwieldly relative weighting for ranking videos with or without transcripts.

Also note that (Tab. 1) prior to the CLIP revolution in video-to-text retrieval, using the audio modality was not uncommon amongst competitive video-to-text retrieval methods (Mithun et al. 2018; Liu et al. 2019a). The trend over the past year, however, has been to instead focus on using only visual features, with *all* recent competitive methods being based off of CLIP, and not using audio data. Our approach, through leveraging commonsense reasoning stored in the LMs, is able to once again allow audio data to enable progress in this common video understanding task, beyond what CLIP alone can provide.

6 Unsupervised Socratic Model Selection

The combination of complementary models, in which one may compensate for the weaknesses of the other, opens an interesting avenue for unsupervised evaluation of model performance. Since our metric of interest is the *combined* performance of e.g., a VLM and a LM – rather than asking the question: '(A): how well does this VLM perform in absolute?', we can instead ask: '(B): how well does this VLM compensate for the weakness of the LM?'.

(Strope et al. 2011) proposes a scheme which does so without requiring any evaluation ground truth. They also find that asking question (B) correlates well with answers to question (A), and is useful e.g., for model selection. The method assumes you have access to a weak (wLM) and a strong (sLM) LM (respectively VLM if evaluating the LM's performance). Asking "how well does this VLM compensate for the weaknesses of the LM" is equivalent to asking: "if we have a collection of VLMs, and we combine them with a weak LM, which model is going to perform the closest to the combination of the VLM with a strong LM?" If a VLM combined with a weak LM, instead of a strong one, makes up for the LM's shortcomings and still performs well in combination, then it may serve as a better component in the context of this combined system.

The benefit of this approach, while not entirely making up for doing absolute evaluations against a ground truth, is that because it only measures relative distance between model outputs, it can be performed unsupervised without annotated data: the distance between the output of the weak and strong combination can be measured using measures of semantic distance, for instance here by scoring them against a distinct, held-out language model.

As an example of using this approach, we extend the method in (Strope et al. 2011) to Socratic Models on egocentric perception, where we show it is possible to quantify the mutual dependence between foundation models without ground truth data. Specifically, to evaluate a new VLM (*VLM'*) for generating language-based world-state history, we first use a baseline VLM *VLM* paired with the strong LM (*sLM*) to generate pseudo ground truth predictions *VLM*×*sLM*. We then take both the baseline VLM *VLM* and new VLM *VLM'*, and pair them with a weak LM *wLM* to generate predictions *VLM*× *wLM* and *VLM'*×*wLM* respectively. We score these predictions (per image summary) against the pseudo ground truth *VLM*×*sLM*. Since the outputs are linguistic, we can measure the similarity of a given prediction to the ground truth, by comparing their sentence embeddings produced by another language model e.g., RoBERTa (Liu et al. 2019b). It is important to use a distinct LM for scoring to avoid spurious correlations with the models under evaluation.

	VLM (CLIP) Variants + Weak LM				
Truth Models	RN50x4	RN50x16	ViT-B/32	ViT-B/16	ViT-L/14
GPT-3 + ViT-B/16	0.628	0.646	0.686	0.861	**0.704**
GPT-3 + RN50x16	0.667	0.851	0.689	0.655	**0.704**
ImageNet Accuracy	65.8	70.5	63.2	68.6	76.2
Size (# params)	178M	291M	151M	150M	427M

Tab. 3: Unsupervised evaluation (higher is better) of various VLMs by pairing them with a weak LM and comparing outputs to a VLM paired with a strong LM, which provides relative 'truth gradients' that inform how well the VLMs can compensate for the weak LM. These results suggest that better VLMs (measured by zero-shot ImageNet classification accuracies) can improve Socratic synergies.

Tab. 3 shows example results of this analysis with GPT-3 "Davinci" as the *sLM*, and GPT-3 "Curie" as the *wLM*, to compare VLM (i.e., CLIP) variants with different backbones: vision transformers (ViT) (Dosovitskiy et al. 2020) and ResNets (RN50) (He et al. 2016) with different model sizes. We find that this method can capture a correlation of ascending performance curve with increasingly better VLMs (e.g., better variants of CLIP) (Radford et al. 2021), as measured by zero-shot image classification accuracy on ImageNet (Deng et al. 2009) – with correlation coefficients of 0.41 and 0.46 between ImageNet accuracies and mean similarity to truth models via ViT-B/16 and RN50x16 respectively. We find that with our SM system for egocentric perception (and in contrast to the original setting in (Strope et al. 2011)), it is necessary to use a third baseline VLM ***b****VLM*×*sLM* to generate the pseudo ground truth, instead of *VLM*×*sLM*. This is because the SM combinations that use the same VLM as the one that generates ground truth are biased to produce similar visual grounding results and can exhibit an unfair advantage during the comparisons. Those numbers in our tests have been grayed out in Tab. 3.

7 Related Work

Multi-model multimodal reasoning. In the context of transfer learning (e.g., via fine-tuning), pre-trained foundation models have achieved strong results when combined and trained together for a number of downstream multimodal (Ngiam et al. 2011) applications including VLMs with LMs for image captioning (e.g., CLIP with GPT-2) (Mokady, Hertz, and Bermano 2021), video understanding (e.g., CLIP with BERT (Gao et al. 2021)), visual question answering e.g., (Song et al. 2022a) and ALMs and LMs for speech and text modeling e.g., (Song et al. 2022b; Bapna et al. 2022). These systems are often finetuned on task-specific data, and while this paradigm is likely to be preferred in domains for which data is abundant, our initial results suggest that SMs can be a strong zero-shot alternative for applications in which data is less available or more expensive to obtain, e.g., egocentric perception and robotics.

The notion as well of "Mixture-of-Experts" ((Jordan and Jacobs 1994), see (Masoudnia and Ebrahimpour 2014) for a review) is a common paradigm for combining the outputs of multiple models, and specifically mixtures of experts across multimodal domains including vision and audio (Liu et al. 2019a) have been studied – note that results from Liu et al. (2019a) are included in Table 1. Investigating further these techniques in the context of recent pretrained foundation models may be a promising direction for future work. Our work may be interpreted as a particular extension of Mixture-of-Experts in which experts may be composed to provide feedback to each other, closed-loop, via the common representation of language.

Egocentric perception continues to be an important problem in computer vision. Early work in the area explores hand-designed first-person visual features for egocentric action recognition, object understanding, and video summarization. This includes ego-motion (e.g., optical flows) (Kitani et al. 2011; Ryoo and Matthies 2013) as well as features from human gaze, hands, and objects (Spriggs, De La Torre, and Hebert 2009; Lee, Ghosh, and Grauman 2012; Fathi, Farhadi, and Rehg 2011; Pirsiavash and Ramanan 2012; Li and Kitani 2013; Lee and Grauman 2015). Focusing on hand-designed features was common in early egocentric vision research, as the availability of data (or videos in general) was very limited. More recent approaches in egocentric perception leverage learned feature representations, utilizing pre-trained convolutional network features (Ryoo, Rothrock, and Matthies 2015), finetuning them (Ma, Fan, and Kitani 2016; Zellers et al. 2022), or training them from scratch (Bambach et al. 2015) with first-person videos. Similar to the topics explored in early work, learning of visual representations capturing human hands, objects, and eye gaze has been extensively studied (Garcia-Hernando et al. 2018; Li, Liu, and Rehg 2018). (Kazakos et al. 2019) learns multimodal embeddings (i.e., video + audio), and (Furnari and Farinella 2019) studies future action anticipation from egocentric videos. Lack of sufficient data however, consistently remains a bottleneck – motivating researchers to construct new larger-scale egocentric video datasets including EPIC-Kitchens (Damen et al. 2018), Charades-Ego (Sigurdsson et al. 2018), and Ego4D (Grauman et al. 2021).

8 Discussion

Socratic Models are a class of systems that leverage structured dialogue between multiple language-based foundation models to make joint predictions for new multimodal tasks. SMs leverage the commonsense knowledge already stored within foundation models pretrained on different domains of data (e.g., text-to-text, text-to-images, text-to-audio), which may include for example Internet-scale data. Our shown systems for egocentric perception, image captioning, and video-to-text retrieval are just examples of the SMs framework, and may shed light on new opportunities to build simple systems that adapt foundation models to (i) capture new multimodal functionalities zero-shot without having to rely on additional domain-specific data collection or model finetuning, and (ii) do so while retaining their robustness to distribution shifts (which is known to deteriorate after finetuning) (Wortsman et al. 2021).

SMs present a language-based approach to combining the outputs of multiple foundation models, which differs from a classical Bayesian approach where one model is used as a prior and the other as evidence. Relying on language-only multi-model discussion carries both pros and cons. For example, the intermediate outputs of the models may be more interpretable, but are treated as "truth" between models – i.e., not weighing them against the other's priors or evidence, which can lead to more divergent model interactions.

In the context of egocentric perception, we find that formulating video Q&A as reading comprehension in SMs directly leverages the extent to which large LMs are capable of logical reasoning by connecting commonsense relationships with knowledge learned from Internet-scale data. For example, the system returns the following answer when presented with the world-state history log:

```
8:00 AM: went to grocery store to buy orange juice, chocolate, and bread.
8:15 AM: I went to gas station to fill up the vehicle tank.
8:30 AM: drove back home and left the groceries in the kitchen.
8:45 AM: started cooking eggs in the pan.
9:00 AM: the dog went into the kitchen.
9:15 AM: took the dog out for a walk.
9:30 AM: the dog is sick.
Q: Why is the dog sick? A: The dog may have eaten something it was not supposed to, such as chocolate.
```

Arriving at the answer requires bridging multiple connections between observations e.g., that the dog went into the kitchen, that the groceries are still in the kitchen, and that the groceries contain chocolate. Such results offer a glimpse of what might be possible using SMs for deductive reasoning across multiple domains of information, and raises interesting research questions on (i) how to better assemble language-based world-state histories (beyond what is presented in this work) that capture relevant evidence to improve the accuracy of conclusions, and (ii) how to elicit chain of thought prompting (Wei et al. 2022) to decompose multi-step problems into intermediate ones. For example, one promising extension could be prompting the LM with chain of thought sequences to expand on hypotheses:

```
Q: What are reasons for why I might be chopping wood?
A: Reasons might include: needing firewood, wanting
to make a statement, or needing the exercise.
```

to which each hypothesis can be progressively explored by downstream subprograms called at recursively higher resolutions until a conclusion is reached. These directions suggest pathways towards achieving increasingly meaningful utility and analysis by digital multimodal assistants.

8.1 Broader Impacts

Socratic Models offer a new perspective that encourages building AI systems using off-the-shelf language-interactable foundation models without additional data collection or model finetuning. This leads to several practical benefits, new applications, and risks as well. For one, SMs provide an interpretable window, through language, into the behavior of the systems (even for non-experts). Further, the barrier to entry for this technology is small: SMs can be engineered to capture new functionalities with minimal compute resources. No model training was used to create any demonstrated results. This can be enabling, but also raises potential risks, since it increases the flexibility of unintended end use applications, and should be carefully monitored over time. We welcome broad discussion on how to maximize the potential positive impacts (enabling broad, new multimodal applications, with minimal new resources) while minimizing the capabilities of bad actors.

Regarding the impact on energy and other resource consumption for machine learning, this work may help pave a path for new, capable machine learning models to be composed with minimal training resource consumption, provided that large foundational pretrained models are available. This may help provide an answer for how large pretrained models may be retargeted to a wide variety of multimodal applications, without additional considerable compute resources required. Since SMs help demonstrate how a wide variety of applications may be addressed with fixed (pretrained) models zero-shot, this may also help foster adoption of new machine learning accelerators (e.g., fixed analog circuity (Reuther et al. 2020), optical diffraction (Lin et al. 2018)) for inference with substantially lower power consumption and more compact form factors.

We are excited about opportunities as well in downstream applications. For example, SMs suggest promising research directions for data-driven learning in robotics, where the various modules within a robot system (e.g., planning (Ahn et al. 2022; Huang et al. 2022), perception (Shridhar, Manuelli, and Fox 2022)) can be replaced with zero-shot foundation models imbued with commonsense priors across domains. These ideas may give rise to a new class of robot systems where by grounding affordances (Zeng 2019) on language, control algorithms can begin to tap into the capabilities of models trained on Internet-scale data, and to tackle applications that have traditionally been data-scarce.

9 Acknowledgements

We thank Debidatta Dwibedi and Matthew O'Kelly for excellent feedback on improving this manuscript, Anelia Angelova, Jean-Jacques Slotine, Jonathan Tompson, Maria Attarian, Shuran Song, for fruitful technical discussions, Kan Huang for applications support, Ahmed Omran, Aren Jensen, Malcolm Slaney, Karolis Misiunas for advice on audio models, and Cody Wanner for YouTube videos.

References

Abuzaid, F.; Sethi, G.; Bailis, P.; and Zaharia, M. 2019. To index or not to index: Optimizing exact maximum inner product search. In *35th IEEE International Conference on Data Engineering, ICDE 2019, Macao, China, April 8-11, 2019*, 1250–1261. IEEE.

Agarwal, P.; Betancourt, A.; Panagiotou, V.; and Díaz-Rodríguez, N. 2020. Egoshots, an ego-vision life-logging dataset and semantic fidelity metric to evaluate diversity in image captioning models. *arXiv preprint arXiv:2003.11743*.

Ahn, M.; Brohan, A.; Brown, N.; Chebotar, Y.; Cortes, O.; David, B.; Finn, C.; Gopalakrishnan, K.; Hausman, K.; Herzog, A.; Hsu, J.; Ibarz, J.; Ichter, B.; Irpan, A.; Jang, E.; Ruano, R. J.; Jeffrey, K.; Jesmonth, S.; Joshi, N.; Julian, R.; Kalashnikov, D.; Kuang, Y.; Lee, K.-H.; Levine, S.; Lu, Y.; Luu, L.; Parada, C.; Pastor, P.; Quiambao, J.; Rao, K.; Rettinghouse, J.; Reyes, D.; Sermanet, P.; Sievers, N.; Tan, C.; Toshev, A.; Vanhoucke, V.; Xia, F.; Xiao, T.; Xu, P.; Xu, S.; and Yan, M. 2022. Do as i can and not as i say: Grounding language in robotic affordances. In *arXiv preprint arXiv:2022.00000*.

Apostolidis, E.; Adamantidou, E.; Metsai, A. I.; Mezaris, V.; and Patras, I. 2021. Video summarization using deep neural networks: A survey. *Proceedings of the IEEE* 109(11):1838–1863.

Bain, M.; Nagrani, A.; Varol, G.; and Zisserman, A. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1728–1738.

Bambach, S.; Lee, S.; Crandall, D. J.; and Yu, C. 2015. Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 1949–1957.

Bapna, A.; Cherry, C.; Zhang, Y.; Jia, Y.; Johnson, M.; Cheng, Y.; Khanuja, S.; Riesa, J.; and Conneau, A. 2022. mslam: Massively multilingual joint pre-training for speech and text. *arXiv preprint arXiv:2202.01374*.

Barbieri, M.; Agnihotri, L.; and Dimitrova, N. 2003. Video summarization: methods and landscape. In *Internet Multimedia Management Systems IV*, volume 5242, 1–13. International Society for Optics and Photonics.

Bommasani, R.; Hudson, D. A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M. S.; Bohg, J.; Bosselut, A.; Brunskill, E.; et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33:1877–1901.

Chen, H.; Xie, W.; Vedaldi, A.; and Zisserman, A. 2020. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 721–725. IEEE.

Chen, M.; Tworek, J.; Jun, H.; Yuan, Q.; Pinto, H. P. d. O.; Kaplan, J.; Edwards, H.; Burda, Y.; Joseph, N.; Brockman, G.; et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Cheng, X.; Lin, H.; Wu, X.; Yang, F.; and Shen, D. 2021. Improving video-text retrieval by multi-stream corpus alignment and dual softmax loss. *arXiv preprint arXiv:2109.04290*.

Choromanski, K.; Chen, H.; Lin, H.; Ma, Y.; Sehanobish, A.; Jain, D.; Ryoo, M. S.; Varley, J.; Zeng, A.; Likhosherstov, V.; Kalashnikov, D.; Sindhwani, V.; and Weller, A. 2021a. Hybrid random features. *to appear in ICLR 2022* abs/2110.04367.

Choromanski, K. M.; Likhosherstov, V.; Dohan, D.; Song, X.; Gane, A.; Sarlós, T.; Hawkins, P.; Davis, J. Q.; Mohiuddin, A.; Kaiser, L.; Belanger, D. B.; Colwell, L. J.; and Weller, A. 2021b. Rethinking attention with performers. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Damen, D.; Doughty, H.; Farinella, G. M.; Fidler, S.; Furnari, A.; Kazakos, E.; Moltisanti, D.; Munro, J.; Perrett, T.; Price, W.; et al. 2018. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 720–736.

Damen, D.; Doughty, H.; Farinella, G. M.; Furnari, A.; Kazakos, E.; Ma, J.; Moltisanti, D.; Munro, J.; Perrett, T.; Price, W.; et al. 2020. Rescaling egocentric vision. *arXiv preprint arXiv:2006.13256*.

Del Molino, A. G.; Tan, C.; Lim, J.-H.; and Tan, A.-H. 2016. Summarization of egocentric videos: A comprehensive survey. *IEEE Transactions on Human-Machine Systems* 47(1):65–76.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Fan, C.; Zhang, Z.; and Crandall, D. J. 2018. Deepdiary: Lifelogging image captioning and summarization. *Journal of Visual Communication and Image Representation* 55:40–55.

Fang, H.; Xiong, P.; Xu, L.; and Chen, Y. 2021. Clip2video: Mastering video-text retrieval via image clip. *arXiv preprint arXiv:2106.11097*.

Fathi, A.; Farhadi, A.; and Rehg, J. M. 2011. Understanding egocentric activities. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 407–414.

Furnari, A., and Farinella, G. M. 2019. What would you expect? anticipating egocentric actions with rolling-unrolling lstms and modality attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 6252–6261.

Gao, Z.; Liu, J.; Chen, S.; Chang, D.; Zhang, H.; and Yuan, J. 2021. Clip2tv: An empirical study on transformer-based methods for video-text retrieval. *arXiv preprint arXiv:2111.05610*.

Garcia-Hernando, G.; Yuan, S.; Baek, S.; and Kim, T.-K. 2018. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 409–419.

Grauman, K.; Westbury, A.; Byrne, E.; Chavis, Z.; Furnari, A.; Girdhar, R.; Hamburger, J.; Jiang, H.; Liu, M.; Liu, X.; et al. 2021. Ego4d: Around the world in 3,000 hours of egocentric video. *arXiv preprint arXiv:2110.07058*.

Gu, X.; Lin, T.-Y.; Kuo, W.; and Cui, Y. 2021. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 2961–2969.

Hoai, M., and De la Torre, F. 2014. Max-margin early event detectors. *International Journal of Computer Vision* 107(2):191–202.

Hu, R., and Singh, A. 2021. Transformer is all you need: Multimodal multitask learning with a unified transformer. *arXiv e-prints* arXiv–2102.

Huang, W.; Abbeel, P.; Pathak, D.; and Mordatch, I. 2022. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. *arXiv preprint arXiv:2201.07207*.

Izadi, S.; Kim, D.; Hilliges, O.; Molyneaux, D.; Newcombe, R.; Kohli, P.; Shotton, J.; Hodges, S.; Freeman, D.; Davison, A.; et al. 2011. Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, 559–568.

Jain, A.; Guo, M.; Srinivasan, K.; Chen, T.; Kudugunta, S.; Jia, C.; Yang, Y.; and Baldridge, J. 2021. Mural: multimodal, multitask retrieval across languages. *arXiv preprint arXiv:2109.05125*.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, 4904–4916. PMLR.

Jordan, M. I., and Jacobs, R. A. 1994. Hierarchical mixtures of experts and the em algorithm. *Neural computation* 6(2):181–214.

Kazakos, E.; Nagrani, A.; Zisserman, A.; and Damen, D. 2019. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 5492–5501.

Kitani, K. M.; Okabe, T.; Sato, Y.; and Sugimoto, A. 2011. Fast unsupervised ego-action learning for first-person sports videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3241–3248.

Kitani, K. M.; Ziebart, B. D.; Bagnell, J. A.; and Hebert, M. 2012. Activity forecasting. In *European Conference on Computer Vision (ECCV)*, 201–214.

Kuznetsova, A.; Rom, H.; Alldrin, N.; Uijlings, J.; Krasin, I.; Pont-Tuset, J.; Kamali, S.; Popov, S.; Malloci, M.; Kolesnikov, A.; et al. 2020. The open images dataset v4. *International Journal of Computer Vision* 128(7):1956–1981.

Lee, Y. J., and Grauman, K. 2015. Predicting important objects for egocentric video summarization. *International Journal of Computer Vision* 114(1):38–55.

Lee, Y. J.; Ghosh, J.; and Grauman, K. 2012. Discovering important people and objects for egocentric video summarization.

In *IEEE conference on computer vision and pattern recognition (CVPR)*, 1346–1353.

Lei, J.; Yu, L.; Berg, T. L.; and Bansal, M. 2020. What is more likely to happen next? video-and-language future event prediction. *arXiv preprint arXiv:2010.07999*.

Li, C., and Kitani, K. M. 2013. Pixel-level hand detection in egocentric videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3570–3577.

Li, J.; Selvaraju, R.; Gotmare, A.; Joty, S.; Xiong, C.; and Hoi, S. C. H. 2021a. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in Neural Information Processing Systems* 34.

Li, Y.; Nagarajan, T.; Xiong, B.; and Grauman, K. 2021b. Ego-exo: Transferring visual representations from third-person to first-person videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6943–6953.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*.

Li, Y.; Liu, M.; and Rehg, J. M. 2018. In the eye of beholder: Joint learning of gaze and actions in first person video. In *Proceedings of the European conference on computer vision (ECCV)*, 619–635.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.

Lin, X.; Rivenson, Y.; Yardimci, N. T.; Veli, M.; Luo, Y.; Jarrahi, M.; and Ozcan, A. 2018. All-optical machine learning using diffractive deep neural networks. *Science* 361(6406):1004–1008.

Liu, Y.; Albanie, S.; Nagrani, A.; and Zisserman, A. 2019a. Use what you have: Video retrieval using representations from collaborative experts. *BMVC*.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Luo, H.; Ji, L.; Zhong, M.; Chen, Y.; Lei, W.; Duan, N.; and Li, T. 2021. Clip4clip: An empirical study of clip for end to end video clip retrieval. *arXiv preprint arXiv:2104.08860*.

Ma, M.; Fan, H.; and Kitani, K. M. 2016. Going deeper into first-person activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1894–1903.

Masoudnia, S., and Ebrahimpour, R. 2014. Mixture of experts: a literature survey. *Artificial Intelligence Review* 42(2):275–293.

Miech, A.; Alayrac, J.-B.; Smaira, L.; Laptev, I.; Sivic, J.; and Zisserman, A. 2020. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9879–9889.

Mithun, N. C.; Li, J.; Metze, F.; and Roy-Chowdhury, A. K. 2018. Learning joint embedding with multimodal cues for cross-modal video-text retrieval. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, 19–27.

Mokady, R.; Hertz, A.; and Bermano, A. H. 2021. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*.

Ngiam, J.; Khosla, A.; Kim, M.; Nam, J.; Lee, H.; and Ng, A. Y. 2011. Multimodal deep learning. In *ICML*.

Oncescu, A.-M.; Koepke, A.; Henriques, J. F.; Akata, Z.; and Albanie, S. 2021. Audio retrieval with natural language queries. *arXiv preprint arXiv:2105.02192*.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C. L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Preprint*.

Patel, D.; Parikh, R.; and Shastri, Y. 2021. Recent advances in video question answering: A review of datasets and methods. In *International Conference on Pattern Recognition*, 339–356. Springer.

Patrick, M.; Huang, P.-Y.; Asano, Y.; Metze, F.; Hauptmann, A.; Henriques, J.; and Vedaldi, A. 2020. Support-set bottlenecks for video-text representation learning. *arXiv preprint arXiv:2010.02824*.

Petroni, F.; Rocktäschel, T.; Lewis, P.; Bakhtin, A.; Wu, Y.; Miller, A. H.; and Riedel, S. 2019. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*.

Pirsiavash, H., and Ramanan, D. 2012. Detecting activities of daily living in first-person camera views. In *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2847–2854.

Portillo-Quintero, J. A.; Ortiz-Bayliss, J. C.; and Terashima-Marín, H. 2021. A straightforward framework for video retrieval using clip. In *Mexican Conference on Pattern Recognition*, 3–12. Springer.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763. PMLR.

Rajpurkar, P.; Jia, R.; and Liang, P. 2018. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*.

Ramsauer, H.; Schäfl, B.; Lehner, J.; Seidl, P.; Widrich, M.; Gruber, L.; Holzleitner, M.; Adler, T.; Kreil, D. P.; Kopp, M. K.; Klambauer, G.; Brandstetter, J.; and Hochreiter, S. 2021. Hopfield networks is all you need. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Rawat, A. S.; Chen, J.; Yu, F. X.; Suresh, A. T.; and Kumar, S. 2019. Sampled softmax with random fourier features. In Wallach, H. M.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E. B.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 13834–13844.

Reuther, A.; Michaleas, P.; Jones, M.; Gadepally, V.; Samsi, S.; and Kepner, J. 2020. Survey of machine learning accelerators. In *2020 IEEE high performance extreme computing conference (HPEC)*, 1–12. IEEE.

Rhinehart, N., and Kitani, K. M. 2017. First-person activity forecasting with online inverse reinforcement learning. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 3696–3705.

Ryoo, M. S., and Matthies, L. 2013. First-person activity recognition: What are they doing to me? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2730–2737.

Ryoo, M. S.; Rothrock, B.; and Matthies, L. 2015. Pooled motion features for first-person videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 896–904.

Ryoo, M. S. 2011. Human activity prediction: Early recognition of ongoing activities from streaming videos. In *Proceedings of the*

IEEE International Conference on Computer Vision (ICCV), 1036–1043.

Sharma, P.; Ding, N.; Goodman, S.; and Soricut, R. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2556–2565.

Shridhar, M.; Manuelli, L.; and Fox, D. 2022. Cliport: What and where pathways for robotic manipulation. In *Conference on Robot Learning*, 894–906. PMLR.

Shrivastava, A., and Li, P. 2014. Asymmetric LSH (ALSH) for sublinear time maximum inner product search (MIPS). In Ghahramani, Z.; Welling, M.; Cortes, C.; Lawrence, N. D.; and Weinberger, K. Q., eds., *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, 2321–2329.

Sigurdsson, G. A.; Gupta, A.; Schmid, C.; Farhadi, A.; and Alahari, K. 2018. Charades-ego: A large-scale dataset of paired third and first person videos. *arXiv preprint arXiv:1804.09626*.

Smaira, L.; Carreira, J.; Noland, E.; Clancy, E.; Wu, A.; and Zisserman, A. 2020. A short note on the kinetics-700-2020 human action dataset. *arXiv preprint arXiv:2010.10864*.

Song, H.; Dong, L.; Zhang, W.-N.; Liu, T.; and Wei, F. 2022a. Clip models are few-shot learners: Empirical studies on vqa and visual entailment. *arXiv preprint arXiv:2203.07190*.

Song, Y.; Fan, X.; Yang, Y.; Ren, G.; and Pan, W. 2022b. Large pretrained models on multimodal sentiment analysis. In *Artificial Intelligence in China*. Springer. 506–513.

Spriggs, E. H.; De La Torre, F.; and Hebert, M. 2009. Temporal segmentation and activity classification from first-person sensing. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 17–24.

Strope, B.; Beeferman, D.; Gruenstein, A.; and Lei, X. 2011. Unsupervised testing strategies for asr.

Tancik, M.; Casser, V.; Yan, X.; Pradhan, S.; Mildenhall, B.; Srinivasan, P. P.; Barron, J. T.; and Kretzschmar, H. 2022. Block-nerf: Scalable large scene neural view synthesis. *arXiv preprint arXiv:2202.05263*.

Thoppilan, R.; De Freitas, D.; Hall, J.; Shazeer, N.; Kulshreshtha, A.; Cheng, H.-T.; Jin, A.; Bos, T.; Baker, L.; Du, Y.; et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Vondrick, C.; Pirsiavash, H.; and Torralba, A. 2016. Anticipating visual representations from unlabeled video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 98–106.

Wang, Z.; Yu, J.; Yu, A. W.; Dai, Z.; Tsvetkov, Y.; and Cao, Y. 2021. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*.

Wang, Q.; Zhang, Y.; Zheng, Y.; Pan, P.; and Hua, X.-S. 2022. Disentangled representation learning for text-video retrieval. *arXiv preprint arXiv:2203.07111*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Chi, E.; Le, Q.; and Zhou, D. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Wortsman, M.; Ilharco, G.; Li, M.; Kim, J. W.; Hajishirzi, H.; Farhadi, A.; Namkoong, H.; and Schmidt, L. 2021. Robust fine-tuning of zero-shot models. *arXiv preprint arXiv:2109.01903*.

Wu, B.; Chen, W.; Fan, Y.; Zhang, Y.; Hou, J.; Liu, J.; and Zhang, T. 2019. Tencent ml-images: A large-scale multi-label image database for visual representation learning. *IEEE Access* 7:172683–172693.

Wu, H.-H.; Seetharaman, P.; Kumar, K.; and Bello, J. P. 2021a. Wav2clip: Learning robust audio representations from clip. *arXiv preprint arXiv:2110.11499*.

Wu, J.; Ouyang, L.; Ziegler, D. M.; Stiennon, N.; Lowe, R.; Leike, J.; and Christiano, P. 2021b. Recursively summarizing books with human feedback. *arXiv preprint arXiv:2109.10862*.

Xu, J.; Mei, T.; Yao, T.; and Rui, Y. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5288–5296.

Xu, H.; Ghosh, G.; Huang, P.-Y.; Okhonko, D.; Aghajanyan, A.; Metze, F.; Zettlemoyer, L.; and Feichtenhofer, C. 2021. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*.

Yang, Z.; Gan, Z.; Wang, J.; Hu, X.; Lu, Y.; Liu, Z.; and Wang, L. 2021. An empirical study of gpt-3 for few-shot knowledge-based vqa. *arXiv preprint arXiv:2109.05014*.

Yu, Y.; Kim, J.; and Kim, G. 2018. A joint sequence fusion model for video question answering and retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 471–487.

Zellers, R.; Lu, J.; Lu, X.; Yu, Y.; Zhao, Y.; Salehi, M.; Kusupati, A.; Hessel, J.; Farhadi, A.; and Choi, Y. 2022. Merlot reserve: Neural script knowledge through vision and language and sound. *arXiv preprint arXiv:2201.02639*.

Zeng, A. 2019. *Learning visual affordances for robotic manipulation*. Ph.D. Dissertation, Princeton University.

Zhai, X.; Wang, X.; Mustafa, B.; Steiner, A.; Keysers, D.; Kolesnikov, A.; and Beyer, L. 2021. Lit: Zero-shot transfer with locked-image text tuning. *arXiv preprint arXiv:2111.07991*.

Zhao, Y.; Hessel, J.; Yu, Y.; Lu, X.; Zellers, R.; and Choi, Y. 2021. Connecting the dots between audio and text without parallel data through visual knowledge transfer. *arXiv preprint arXiv:2112.08995*.

Zhou, B.; Khosla, A.; Lapedriza, A.; Torralba, A.; and Oliva, A. 2016. Places: An image database for deep scene understanding. *arXiv preprint arXiv:1610.02055*.