arXiv:2204.00598v2 [cs.CV] 27 May 2022

Socratic Models: Composing Zero-Shot Multimodal Reasoning with Language

Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Choromanski, Adrian Wong,
Stefan Welker, Federico Tombari, Aveek Purohit, Michael S. Ryoo,
Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, Pete Florence

Google

Abstract

Large pretrained (e.g., "foundation") models exhibit distinct 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., spreadsheets, SAT questions, code). As a result, these models store different forms of commonsense knowledge across different domains. In this work, we show that this diversity is symbiotic, and can be leveraged through Socratic Models (SMs): a modular framework in which multiple pretrained models may be composed zero-shot i.e., via multimodal-informed prompting, to exchange information with each other and capture new multimodal capabilities, without requiring finetuning. With minimal engineering, SMs are not only competitive with state-of-the-art zero-shot image captioning and video-to-text retrieval, but also enable new applications such as (i) answering free-form questions about egocentric video, (ii) engaging in multimodal assistive dialogue with people (e.g., for cooking recipes) by interfacing with external APIs and databases (e.g., web search), and (iii) robot perception and planning.

1 Introduction

Large pretrained models (e.g., BERT [1], GPT-3 [2], CLIP [3]) have enabled impressive capabilities [4]: from zero-shot image classification [3, 5], to high-level planning [6, 7]. Their capabilities depend on their training data – while they may be broadly crawled from the web, their distributions remain distinct across domains. For example, in terms of linguistic data, visual-language models (VLMs) [8, 9] are trained on image and video captions, but large language models (LMs) [1, 10, 11] 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., reading comprehension [12]). In this work, we propose these model differences are complementary and can be jointly leveraged to compose (via prompting) new multimodal capabilities out-of-the-box. To this end, we present Socratic Models[1] (SMs), a modular framework in which new tasks are formulated as a language-based exchange between pretrained models and other modules, without additional training or finetuning. These modules can either contain (i) large pretrained ("foundation" [4]) models, or (ii) APIs that interface with external capabilities or databases (e.g., web search, robot actions). Rather than scaling task-specific multimodal training data in the areas of overlap (e.g., alt-text captions [13]), or unifying model architectures for multitask learning [14], SMs embrace the zero-shot capabilities of pretrained models by prompt engineering guided multimodal discussions between the independent models to perform joint inference on a task-specific output.

---

[1] The name draws from an analogy to the Socratic Method, but between modules interacting through language.

Preprint. Under review.

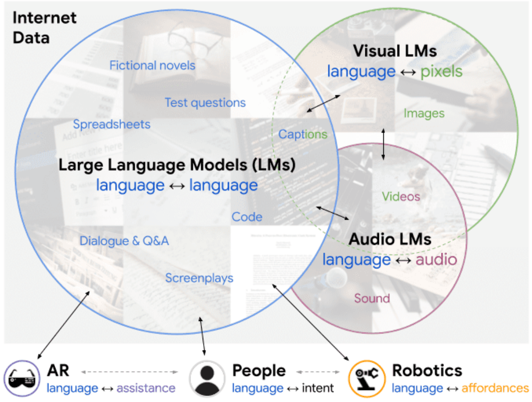


Figure 1: Large pretrained "foundation" models trained across different domains learn complementary forms of commonsense, and language is an intermediate representation by which these models can communicate with each other to generate joint predictions for new multimodal tasks, without requiring finetuning. New applications (e.g., augmented reality (AR), human feedback, robotics) can be viewed as adding participants to the multi-model discussion.

Across a number of domains spanning vision, language, and audio modalities – and via a small amount of creative prompt-enabled multimodal composition – SMs are quantitatively competitive with zero-shot state-of-the-art on standard benchmarks including (i) image captioning on MS COCO [15, 16], (ii) contextual image captioning and description (improving 11.3 to 38.9 captioning CIDEr on Concadia [17]), and (iii) video understanding with video-to-text retrieval (from 40.7 to 44.7 zero-shot R@1 on MSR-VTT [18]). SMs also enable new capabilities across applications such as (i) open-ended reasoning for egocentric perception (Fig. 4), (ii) multimodal assistive dialogue to guide a user through a cooking recipe, and (iii) robot perception-driven planning for sequential pick and place. SMs give rise to new opportunities to address classically challenging problems in one domain, by reformulating it as a problem in another. For example, answering free-form questions about first-person videos (e.g., *"why did I go to the front porch today?"*) was previously thought to be out-of-reach for egocentric perception without domain-specific data collection [19, 20]. We show that this is possible with SMs by assembling video into a *language-based world-state history* (in the form of a short story, or event log), then performing various types of open-ended text-prompted tasks (e.g., answering questions) about that world-state history – i.e., formulating video understanding as a reading comprehension problem, for which modern LMs are proficient.

The goal of this paper is (1) to discuss new perspectives on building AI systems that embrace the heterogeneity of pretrained models through structured Socratic dialogue, and (2) to give example demonstrations of what is already possible today with SMs on challenging multimodal tasks. Our primary contribution is (i) the Socratic Models framework, which proposes to compose multimodal pretrained models through language, without requiring training. The SMs framework contains key, enabling components such as the demonstrated (ii) multimodal prompting methods, including (iii) language-based world-state history for video understanding. Additional contributions include (iv) demonstrating strong quantitative performance of example SM systems, setting new zero-shot state-of-the-art on multiple tasks, including in image captioning and video understanding, and (v) providing additional application examples on open-ended egocentric perception, multimodal assistants, and robot perception and planning. Our demonstrated SM systems are not without limitations – we discuss the unreliability inherited from the models on which they are constructed, together with other potential broader impacts (Sec. 6). Code is available at socraticmodels.github.io.

2 Problem Setting, Background, and Related Work

Problem setting. We are interested in creating a variety of multimodal [21] applications enabled by large pretrained models, which can be viewed as a form of transfer [22, 23]: "knowledge" learned from a set of *surrogate tasks* (e.g., text completion, image-text similarity) is applied to new downstream *target tasks* (e.g., image captioning, robot planning). Consider a set of target tasks where each task i seeks a desired map $f^i : \mathcal{X}^i \rightarrow \mathcal{Y}^i$. We are particularly interested in cases where: (i) each input $\mathcal{X}^i$ and/or output $\mathcal{Y}^i$ may contain multiple modalities e.g., from the power set of {language, vision, audio, robot actions}; (ii) there may be many such tasks; (iii) each target task may have little or no training data available; and (iv) models pretrained on the surrogate tasks are available.

Pretraining weights is a dominant paradigm for transfer learning with deep models, in which pretrained model weights (from surrogate tasks) are used to initialize some subset of parameters in the model for the target task, which are then either (a) left frozen, or (b) finetuned. Pretraining deep models has been studied extensively in the unsupervised setting [24, 25, 26, 27, 28], and in the supervised setting was perhaps most popularized by ImageNet [29] pretraining [30, 31, 32, 33], Various forms of pretraining have been ubiquitous in NLP [34, 35, 36, 37, 38, 1, 2]. For each target task, model architectures and/or training procedures may need to be developed that are composed of these pretrained parameters, for which domain expertise may be advantageous. In multimodal training, it is common to leave sub-portions of models, for example ones associated with one but not other modalities, frozen for downstream tasks [39, 40, 41, 42, 43].

Joint training of all modalities on specific target tasks is a common approach to multimodal learning [42, 44, 45, 46, 47, 48]. For each task i one may obtain a large multimodal dataset and train a task-specific map $f^i_{\theta_i}$ with parameters θ_i, some of which may come from pretrained weights, either frozen or finetuned. A benefit of this approach is that it follows the playbook of: (1) curate a big dataset, (2) train a big model, which given enough data and compute has proven to be formidable [49].

Combining both weights from large pretrained models with multimodal joint training, several works have achieved strong results for a number of downstream multimodal applications including VLMs with LMs for image captioning (e.g., CLIP with GPT-2) [45], video understanding (e.g., CLIP with BERT [46]), visual question answering e.g., [47] and ALMs and LMs for speech and text modeling e.g., [50, 51]. 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 results suggest that SMs can be a strong alternative for applications in which data is less available or more expensive to obtain.

Multimodal probabilistic inference is an alternative e.g., Bayesian approach where one model is used as a prior and the other as evidence – with which models from different modalities may perform joint inference [52, 7]. One prominent example is in automatic speech recognition: different language models can be trained separately, then transfer knowledge to a speech-to-text system via priors [52].

The notion of "Mixture-of-Experts" ([53], see [54] for a review) is also a common paradigm for combining the outputs of multiple models – specifically, mixtures of experts across multimodal domains including vision and audio [55] have been studied. Further investigating these techniques in the context of recent pretrained foundation models may be a promising direction for future work.

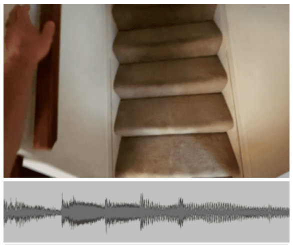
Zero-shot or few-shot prompting recently has been shown, notably by Brown et al. [2], to be highly effective for transfer learning. In this approach, a large pretrained language model is zero-shot or few-shot *prompted* with several examples, without training, to perform a new task. Further methods such as chain-of-thought prompting [56] have shown that even simple prompting modifications can have a profound impact on target task performance [56, 57] and enable new capabilities. Our work builds on these works, by extending prompting methods into the multimodal domain.

3 Socratic Models

Socratic Models (SMs) is a framework in which multiple large pretrained models may be composed through language (via prompting) without requiring training, to perform new downstream multimodal tasks. This offers an alternative method for composing pretrained models that directly uses language as the intermediate representation by which the modules exchange information with each other. It is both distinct from, and may be complementary to, other multimodal approaches such as joint multimodal training (Sec. 2). SMs are perhaps most intuitively understood through examples, which are provided in Sec. 4 and 5, but a definition is as follows. A task-specific Socratic Model $f_{\text{SM}} : \mathcal{X} \rightarrow \mathcal{Y}$ may be described as a computation graph, with nodes as a set of modules $\{f^i_{\mathcal{M}^i}\}$, and the edges of the graph represent intermodule communication through language. Each $\mathcal{M}$ is some (multimodal) model or external API, and each module f assists in transforming the output of one f into a form of language that a connected f' may use for further inference. For visualization, outputs from LMs are blue, VLMs green, ALMs purple, prompt text gray, user inputs magenta, VLM-chosen LM outputs green-underlined blue, and ALM-chosen LM outputs purple-underlined blue.

A key component in SMs is *multi-model multimodal prompting*, in which information from a non-language domain is substituted into a language prompt, which can be used by an LM for reasoning. One way to multimodal prompt is to variable-substitute language-described entities from other modalities into a prompt. An example of this is: $\text{activity} = f_{\text{LM}}(f_{\text{VLM}}(f_{\text{LM}}(f_{\text{ALM}}(f_{\text{LM}}(f_{\text{VLM}}(\text{video}))))))$ shown in Fig. 2, where (i) the VLM detects visual entities, (ii) the LM suggests sounds that may be heard, (iii) the ALM chooses the most likely sound, (iv) the LM suggests possible activities, (v) the VLM ranks the most likely activity, (vi) the LM generates a summary of the Socratic interaction. Some form of such multimodal prompting is central to all of our demonstrated SM examples (Sec. 4

and 5). Note that this example involves multiple back-and-forth interactions, including calling the same model multiple times, forming a sort of “closed-loop” feedback between nodes in the SM graph.



```
I am in a: staircase. I see
a: stairs, animal, mammal,
hamster, human leg. I think
I hear footsteps. I am:
climbing. Summary: I am most
likely climbing a staircase,
and I may hear footsteps.
```

Figure 2: Example: SM systems can be prompted to zero-shot annotate an egocentric image with a summary of the person’s activities. Information from multiple modalities (language, audio) can help denoise predictions from any one specific modality (vision).

Informally SMs may be interpreted as composing pretrained models to “talk to each other”, but in practice certain models may need simple pre- and post-processing to produce language. For example, vision-text similarity VLMs, e.g., CLIP [3], do not inherently produce text, but can be made to perform zero-shot detection from a large pre-existing library of class category names, and return the top-k detected categories. Accordingly, although our example SM systems required no training, the interactions between models are scripted with prompt templates. While in future work we are excited to explore learning the interactions (i.e., forms of each f, and edges), we also find practical benefits of a framework with no required task-specific training: new applications can be quickly targeted by just a small amount of creative programming.

SMs are in part a reaction to the constraints of the predominant “pre-training weights” (Sec. 2) paradigm to transfer learning with foundation models, which include: (i) expensive (at times prohibitively) to finetune large 100B+ parameter models both in terms of compute costs and data collection (can be challenging for new multimodal applications e.g., in AR or robotics), (ii) finetuning pretrained model weights may lose generality and robustness to distribution shifts [58], (iii) foundation models may store “stale” knowledge (due to training latencies), and lack access to dynamic online data or proprietary sources of information. Despite these limitations, large pretrained foundation models [4] are likely to serve as a backbone for many intelligent systems of the future – SMs is a systems approach (i.e., glue framework) that leans on their zero-shot and few-shot capabilities in aggregate as a means to address these limitations for new downstream multimodal tasks.

4 Evaluation: Methods and Results

In this section, we quantitatively evaluate Socratic Models on: image captioning [15, 16] (Sec. 3.1), contextual image captioning [17] (Sec. 3.2), and video-to-text retrieval [18] (Sec. 3.3). For each task, we (i) describe how we use the SMs framework, and (ii) discuss results.

4.1 Socratic Image Captioning on MS COCO Captions: VLM + LM

```
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:
```

SM (ours): This image shows an inviting dining space with plenty of natural light.

ClipCap: A wooden table sitting in front of a window.

SM (ours): People gather under a blossoming cherry tree, enjoying the beauty of nature together.

ClipCap: Students enjoying the cherry blossoms.

SM (ours): At the outdoor market, you can find everything from plantains to Japanese bananas.

ClipCap: A bunch of bananas sitting on top of a table.

Figure 3: SMs with VLM and LM prompting (left) can zero-shot generate captions for generic Internet images (e.g., from MS COCO), and can be as expressive as task-specific finetuned methods such as ClipCap [45].

Method. SMs can generate image captions by prompting a guided language-based exchange between a VLM and LM – i.e., via caption $= f^3_{\text{VLM}}(f^2_{\text{LM}}(f^1_{\text{VLM}}(\text{image})))$. First (1), the VLM is used to zero-shot detect different place categories (Places356 [59]), object categories (from Tencent ML-Images [60]), image type ({*photo, cartoon, sketch, painting}*) and the number of people {*are no people, is one person, are two people, are three people, are several people, are many people*}. The top-k ranked in each category can then be substituted into an LM prompt, as shown in Fig. 3, left. Second (2), given the VLM-informed language prompt, a causal LM (i.e., for text completion) generates several n candidate captions. For this step, we use a non-zero next-token sampling temperature (e.g., 0.9 for

GPT-3), to return sufficiently diverse, but reasonable results across the n candidates. Finally (3), these n captions are then ranked by the VLM with the image, and the highest scoring caption is returned.

Results. Tab. 1 shows quantitative comparisons with state-of-the-art image captioning methods on MS COCO Captions dataset [15, 16]. We chose to evaluate over a random sampled subset of 100 images from the test split [63], so that GPT-3 API runtime costs are more affordable for reproducibility (∼$150 USD per run with with $n = 20$ generated candidate captions per image). Metrics from baselines are comparable to full-test-set metrics (see Appendix).

Method	BLEU-4	METEOR	CIDEr	SPICE	ROUGE-L
*ClipCap [45]	40.7	30.4	152.4	25.2	60.9
†MAGIC [61]	11.4	16.4	56.2	11.3	39.0
ZeroCap [62]	0.0	8.8	18.0	5.6	18.3
SMs 0-shot (ours)	6.9	15.0	44.5	10.1	34.1
SMs 3-shot (ours)	**18.3**	**18.8**	**76.3**	**14.8**	**43.7**

*finetuned on full training set with image-text pairs.
†finetuned on unpaired training set, zero-shot on image-text pairs.

Table 1: Image captioning comparisons on a random subset of $N = 100$ MS COCO test examples.

SMs substantially outperform the zero-shot state-of-the-art ZeroCap [62] with a CIDEr [64] score 18.0 → 44.5, but do not perform as well as methods such as ClipCap [45] which are directly finetuned on the training set. SMs tend to generate verbose and descriptive captions (see qualitative examples in Fig. 3), but may naturally score lower on captioning metrics if they do not match the dataset's distribution of caption labels. This performance gap narrows as SMs are few-shot prompted with 3 random captions from the training set, bringing CIDEr scores up to 76.3, exceeding the performance of MAGIC [61] which finetunes the text generator on the training set's unpaired captions.

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 captioning work with SMs may explore open-vocabulary object detectors [65, 66] to recover more salient details, or combine the outputs of multiple task-specific image captioning models with LMs to assemble a single rich and coherent caption.

4.2 Socratic Contextual Image Description on Concadia: VLM + LM

Method. Concadia [17] is a dataset for contextual image captioning and description, conditioned on the input image and an associated paragraph of article text. In particular, image descriptions describe the visual content in the image (e.g., "a portrait of a man with a beard in a suit") commonly used for accessibility, while image captions link images to article text (e.g., "a photo of Abraham Lincoln"). We evaluate SMs on both tasks, using a similar method to MS COCO captions (Sec. 4.1) but with article-text prompt-substitution (below), and no need for VLM re-ranking. $f^2_{\text{LM}}(f^1_{\text{VLM}}(\text{image}), \text{context})$:

```
I am an intelligent image captioning bot. The article is about: "{article_text}". In this image, I think I see
a {object1}, {object2}, {object3},... A short caption for this image is:
```

Results. We evaluate SMs on the test split of Concadia with 9,691 images (shown in Tab. 2). Despite being zero-shot, SMs outperform the task-specific prior best method, Kreiss et al. [17], that directly finetunes on the training set of 77,534 images, with a CIDEr score improvement 11.3 → 38.9 for generated image captions, and 17.4 → 22.6 for generated image descriptions. We also report numbers for captioning generation conditioned on the image, article text, and ground truth description. This achieves a CIDEr score of 93.8 and suggests the upper bound of performance if SMs are used with VLMs that can produce accurate image descriptions. We also discuss interesting additional findings in the appendix, e.g., that LMs can perform comparably on contextual image captioning even without input images (i.e., only article text as input), which either (i) reflects a strong correlation between the distributions of captions and article texts, and/or (ii) indicates LM training set overlap. Overall, the results on Concadia are promising and suggest that SMs can be used to automatically generate descriptive texts that improve the accessibility of visual content on the web for the low vision community.

Method	Caption Generation	Description Generation
Kreiss et al. [17]	11.3	17.4
SMs (ours)	**38.9**	**22.6**
SMs w/ description	93.8	–

Table 2: SMs zero-shot are competitive on contextual image captioning and description (measured with CIDEr) on the Concadia dataset, outperforming task-specific methods e.g., Kreiss et al. [17] which finetunes on the training set.

4.3 Socratic Video-to-Text Retrieval: VLM + LM + ALM

Method. Socratic Models can be adapted for video-to-text retrieval, a video understanding task commonly benchmarked on MSR-VTT [18]. Our approach leverages commonsense information from audio and language domains to augment the vision-based Portillo-Quintero et al. [67], which computes a similarity measure between the average VLM (i.e., CLIP) features of all video frames per video, and the VLM text features of captions – used to execute video-to-text retrieval with one-to-many nearest neighbor matching. Specifically, our system transcribes audio from the video with speech-to-text ALMs [51] for automatic speech recognition (ASR e.g., via Google Cloud speech-to-text API [68]), then summarizes the transcripts with an LM 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 compute similarity scores of the generated summary to the set of captions with a masked LM (e.g., similarity between sentence embeddings from RoBERTa [69]), and use those scores to re-weight the CLIP-based ranking from Portillo-Quintero et al. For videos with sufficiently-long transcripts ($\geq$100 characters), the matching score is: $\big(\textit{CLIP}\ (\text{caption}) \cdot \textit{CLIP}\ (\text{video}')\big) \times \big(\textit{RoBERTa}\ (\text{caption}) \cdot \textit{RoBERTa}\ (\textit{GPT-3}(\textit{prompt},\ \textit{Speech2Text}\ (\text{audio}')))\big)$, 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 default to Portillo-Quintero et al., which is just $\textit{CLIP}(\text{caption}) \cdot \textit{CLIP}(\text{video}')$. Here, the Socratic interaction lies mainly between the ALM (speech-to-text) to the commonsense LM (GPT-3 to summarize transcriptions), and between the commonsense LM to the ranking based system that is a combination of the VLM (CLIP) and the masked LM (RoBERTa).

Results. We evaluate on MSR-VTT [18], which as noted in other recent works [46, 73] is the most popular benchmark for video-to-text retrieval. We compare our method with both zero-shot methods, as well as finetuned methods specifically trained on MSR-VTT. Results show that our method sets a new zero-shot state-of-the-art (Tab.3). Since our system uses Portillo-Quintero et al. [67] to process CLIP features but additionally incorporates LM reasoning on speech-to-text transcripts, the increased measured performance of our method (i.e., 40.3 → 44.7 R@1) directly reflects the added benefits of incorporating language-based multimodal reasoning. Additionally, to keep the comparison between our method and Portillo-Quintero et al. [67] as direct as possible, we maintain the usage of their precomputed CLIP features from ViT-B/32, but it is likely that performance can be improved with other recent more performant VLMs (e.g., LiT [39], CLIP with ViT-L/14).

		MSR-VTT Full				
Category	Method	R@1↑	R@5↑	R@10↑	MdR↓	Audio
Finetuned	JEMC [70]	12.5	32.1	42.4	16.0	yes
	Collab. Experts [55]	15.6	40.9	55.2	8.3	yes
	CLIP2Video [71]	**54.6**	**82.1**	**90.8**	**1.0**	no
Zero-shot	CLIP via [67]	40.3	69.7	79.2	**2.0**	no
	SMs (ours)	**44.7**	**71.2**	**80.0**	**2.0**	yes

Table 3: Video-to-text retrieval results on MSR-VTT [18] dataset, both on the popular 1k-A [72] 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 [3] is used, which CLIP encoder is used.

Table 4 shows that on the subset of test videos that contain *long-transcripts*, we observe a more substantial increase in performance from 40.3 to 54.9 with our method compared to Portillo-Quintero et al. [67]. Note that this is roughly comparable to the R@1 of the best *finetuned*-SOTA method, CLIP2Video [71], with 54.6 R@1 (Tab. 3). 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 speech present in the audio, our zero-shot SMs can nearly match the *finetuned*-SOTA methods for video-to-text retrieval.

	Long-transcript subset of MSR-VTT Full			
	R@1↑	R@5↑	R@10↑	MdR↓
CLIP via [67]	41.5	69.6	77.4	2.0
SMs (ours)	**54.9**	**74.0**	**79.9**	**1.0**

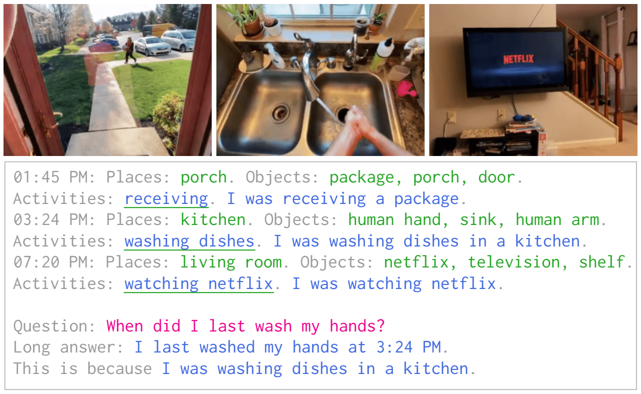
Table 4: SMs substantially improve on Portillo-Quintero et al. [67] for video-to-text retrieval on the MSR-VTT subset of videos for which *long-transcripts* are available (n=1,007 out of 2,990).

5 Applications: Methods and Demonstrations

In this section, we describe several applications of SMs on (i) egocentric perception, (ii) multimodal assistive dialogue, and (iii) robot perception and planning. These applications each involve processing user inputs/feedback, and serve as examples of integrating external modules (e.g., web search, robot policies) as additional participants to a Socratic discussion to enable new multimodal functionalities.

5.1 Egocentric Perception: User + VLM + LM + ALM

SMs can be prompted to perform various perceptual tasks on egocentric video: (i) summarizing content, (ii) answering free-form reasoning questions, (iii) and forecasting. Egocentric perception has downstream applications in AR and robotics, but remains challenging: the characteristics of first-person footage – from unusual viewpoints to lack of temporal curation – are not often found in existing datasets, which focus more on generic Internet content captured from third-person views [29, 16, 74]. This domain shift makes it difficult for data-driven egocentric models to benefit from the paradigm of pretraining on third person Internet data [75, 76]. SMs offer a zero-shot alternative to perform egocentric perceptual tasks without training on large domain-specific datasets [19, 20, 76].



```
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.
07:20 PM: Places: living room. Objects: netflix, television, shelf.
Activities: watching netflix. I was watching netflix.

Question: When did I last wash my hands?
Long answer: I last washed my hands at 3:24 PM.
This is because I was washing dishes in a kitchen.
```

Figure 4: SMs with VLM, LM, and ALM can be prompted to generate a captions for key moments in videos, which can be assembled into a language-based world-state history (e.g., in the form of an event log) that the LM can answer free-form questions about.

For open-ended reasoning, a key aspect of our SMs-based approach is formulating *video understanding as reading comprehension*, i.e., re-framing "video Q&A" as a "short story Q&A" problem, which differs from common paradigms for video understanding that may involve supervising video-text models on labeled datasets or adversarial training (see [77] for a recent survey). To this end, we first extract a set of "key moments" throughout the video (e.g., via importance sampling, or video/audio search based on the input query, discussed in Appendix). We then caption the key frames indexed by these moments (using prompts similar to those in Sec. 4.1 and Sec. 4.2), and recursively summarize [78] them into a language-based record of events, which we term a *language-based world-state history*. This is then passed as context to an LM to perform various reasoning tasks via text completion such as Q&A, for which LMs have demonstrated strong zero-shot performance [2]. Drawing analogies to 3D vision and robotics, the world-state history can be thought of as building an on-the-fly reconstruction of events in the observable world with language, rather than other representations, such as dynamically-updated 3D meshes [79] or neural fields [80].

(i) Summarization enables augmenting human memory to recall events or life-log activities. Given world-state history constructed from SMs using a first-person POV video[2], this can be implemented by prompting an LM to complete: "{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."

(ii) Open-ended Q&A involves prompting the LM to complete the template: "{world-state history} Q: {question} A:". Conditioned on the quality (comprehensiveness) of the world-state history, LMs can generate surprisingly meaningful results to contextual recall questions (e.g., "what was I doing outdoors?" → "I was chopping wood in a yard.", "did I drive today?" → "no, I did not drive today."), temporal questions (e.g., "when did I last drink coffee?" → "I last drank coffee at 10:17 AM", "how many times did I receive a package today?" → "I received a package once today."), cause-and-effect questions (e.g., "why did I go to the front porch today?" → "I went to the front porch today to receive a package."). As in [81] we can also further prompt the LM to *explain the answer* by adding "This is because:" to which it can respond "I saw on the porch a package and knew that I was expecting it."

(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 (example below on the right) are generative, and are more broad than binary event classification [82].

---

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

Few-shot prompting the LM with additional examples of prior event logs most similar to the current one is likely to improve the precision of the results, which may be useful for assistive AR applications. Without additional context, the completions are likely biased towards typical schedules seen by the LM across Internet-scale data.

```
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.
```

5.2 Multimodal Assistive Dialogue: User + VLM + LM + Web Search

SMs can be adapted to engage in multimodal dialogue to assist people in doing every day tasks, such as cooking. Our example application here helps the user search for a recipe, then guides them through it step by step. The system allows the user to navigate recipe steps with casual dialogue, provides ingredient replacements or advice (using LM priors), and searches for visual references (in the form of images or videos) on user request. This is a case study in (i) prompting a dialogue LM [10] to produce key phrase tokens that elicit specific Socratic interactions (e.g., video search via a VLM to output visual data), and (ii) using a web crawler (outputs in red) as an additional module engaged in Socratic discussion with other models to retrieve information online. The approach preconditions an LM (e.g., GPT-3 [2]) with context that includes when and how key phrases should be referenced:

```
Alice is a an expert chef that will help Bob prepare a given recipe. If Bob asks for the next step, Alice will
respond with "Step: " followed by the next step of the recipe. If Bob does not have the right ingredients, Alice
will assist Bob in finding suitable replacements. If Bob asks Alice to describe something that is better shown
visually, Alice will say "(image)" followed by a response. Ingredients: {ingredients} Recipe: {recipe}
```

where phrases such as "(image)" initiate a Socratic interaction where the web crawler returns a video of the recipe (if it exists), and the VLM runs visual search on it using the LM response to retrieve a related frame. We find that this can provide qualitatively meaningful guidance to a user, for example[3]:

```
Alice: Step: In a medium bowl, beat the egg whites until frothy.
Bob: how many eggs.
Alice: 3 egg whites, room temperature
Bob: whats room temperature.
Alice: Room temperature is around 70 degrees Fahrenheit.
Bob: what do you mean by medium bowl.
Alice: A medium bowl is a bowl that is big enough to fit the ingredients for this
recipe and that is comfortable to mix in.
Bob: exactly how big should that be.
Alice: (image) A medium bowl is a bowl that is big enough to fit the ingredients for
this recipe and that is comfortable to mix in.
(a) video search {recipe video} with "A medium bowl is a bowl that is..."
Bob: ok assume im done whats next.
Alice: Step: Keep beating and slowly add the sugar until stiff peaks form.
Bob: whats stiff peaks.
Alice: Stiff peaks are when the egg whites stand up in peaks that do not fall over.
Bob: can you show me what that looks like.
Alice: (image) Stiff peaks are when the egg whites stand up in peaks that do not fall
over.
(b) video search {recipe video} with "Stiff peaks are when the egg whites..."
Bob: do i need to add water.
Alice: No, you do not need to add water.
```

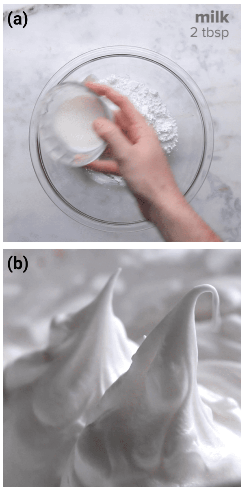


Figure 5: SMs with VLM, Web Search, and LM prompting can enable multimodal dialogue applications such as guiding a user through online recipe steps and providing assistive visuals via video search.

5.3 Robot Perception & Planning: User + VLM + LM + Policies

SMs can be used to enable robots to perform language-conditioned tasks. Our example system uses a VLM (open-vocabulary object detection with ViLD [65]) to describe the objects in the scene, feeds that description as context to a LM as a multi-step planner [7, 6], that then generates the individual steps to be passed to a pretrained language-conditioned robot policy (e.g., models similar to CLIPort [83, 84] for open vocabulary pick-and-place). Steps can be represented in the form of natural language ("Pick the red block and place it on the blue block.") or in the form of pseudocode (to generate text with a fixed template e.g., "robot.pick_and_place("red block", "blue block")"), leveraging LM capacity to write code. We demonstrate this in the context of a simulated environment (shown in Fig.

---

[3] Example using recipe steps and ingredients from tasty.co/recipe/strawberry-cheesecake-macarons

6) using a UR5 arm and and several objects (blocks, bowls). Distinct from [7], this uses cross-modal VLM-informed prompting and LM code generation, rather than joint probabilistic inference.

```
objects = ["green block", "blue block", "yellow block", "green
bowl", "blue bowl", "yellow bowl"]
# move all the blocks to different corners.
Step 1. robot.pick_and_place("green block", "top left corner")
Step 2. robot.pick_and_place("blue block", "top right corner")
Step 3. robot.pick_and_place("yellow block", "bottom left corner")
# now move the blue block to the middle.
Step 1. robot.pick_and_place("blue block", "middle")
# stack the blocks on top of each other.
Step 1. robot.pick_and_place("yellow block", "blue block")
Step 2. robot.pick_and_place("green block", "yellow block")
# wait actually undo that last step.
Step 1. robot.pick_and_place("green block", "top left corner")
# put the yellow block in the bowl you think it best fits.
Step 1. robot.pick_and_place("yellow block", "yellow bowl")
# ok now sort the remaining blocks in the same way.
Step 1. robot.pick_and_place("blue block", "blue bowl")
Step 2. robot.pick_and_place("green block", "green bowl")
```

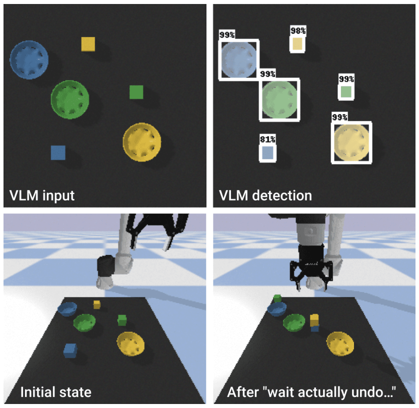


Figure 6: SMs can be engineered with VLM, LM, and language-conditioned robot policies (e.g., via CLIPort [83]) to enable robots to parse and generate plans from free-form human instructions (in magenta).

Chaining this system together expands the set of language-specified tasks beyond the original set of primitives trained by the policy, and enables applications involving human dialogue with the robot.

6 Discussion

Socratic Models is a modular framework that leverages structured dialogue (i.e., via prompting) between multiple large pretrained 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 Internet-scale data. Our shown systems for image captioning, video-to-text retrieval, egocentric perception, multimodal dialogue, robot perception and planning are just examples of the SMs framework, and may shed light on new opportunities to build simple systems that adapt pre-existing 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) [58]. Potential future work may involve meta-learning the Socratic interactions themselves, and extending the inter-module edges to include additional modalities beyond language, e.g., passing images between modules.

Broader Impacts. SMs offer new perspectives that encourage building AI systems using off-the-shelf large pretrained 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 of entry for this technology is small: SMs can be engineered to capture new functionalities with minimal compute resources, and to tackle applications that have traditionally been data-scarce. No model training was used to create our 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. It is also important to note that the system may generate results that reflect unwanted biases found in the Internet-scale data on which incorporated models are trained, and should be used with caution (and checked for correctness) in downstream applications. 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.

Acknowledgments and Disclosure of Funding

We thank Debidatta Dwibedi, Matthew O'Kelly, and Kevin Zakka for excellent feedback on improving this manuscript, Anelia Angelova, Jean-Jacques Slotine, Jonathan Tompson, 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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[15] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

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

[17] E. Kreiss, N. D. Goodman, and C. Potts. Concadia: Tackling image accessibility with context. *arXiv preprint arXiv:2104.08376*, 2021.

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

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

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

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

[22] R. Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

[23] S. Thrun. Lifelong learning algorithms. In *Learning to learn*, pages 181–209. Springer, 1998.

[24] G. E. Hinton, S. Osindero, and Y.-W. Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

[25] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle. Greedy layer-wise training of deep networks. *Advances in neural information processing systems*, 19, 2006.

[26] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008.

[27] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng. Self-taught learning: transfer learning from unlabeled data. In *Proceedings of the 24th international conference on Machine learning*, pages 759–766, 2007.

[28] G. Mesnil, Y. Dauphin, X. Glorot, S. Rifai, Y. Bengio, I. Goodfellow, E. Lavoie, X. Muller, G. Desjardins, D. Warde-Farley, et al. Unsupervised and transfer learning challenge: a deep learning approach. In *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, pages 97–110. JMLR Workshop and Conference Proceedings, 2012.

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

[30] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[31] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning*, pages 647–655. PMLR, 2014.

[32] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

[33] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*, 2013.

[34] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.

[35] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[36] A. M. Dai and Q. V. Le. Semi-supervised sequence learning. *Advances in neural information processing systems*, 28, 2015.

[37] P. Ramachandran, P. J. Liu, and Q. V. Le. Unsupervised pretraining for sequence to sequence learning. *arXiv preprint arXiv:1611.02683*, 2016.

[38] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. 2018.

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

[40] T. D. Kulkarni, A. Gupta, C. Ionescu, S. Borgeaud, M. Reynolds, A. Zisserman, and V. Mnih. Unsupervised learning of object keypoints for perception and control. *Advances in neural information processing systems*, 32, 2019.

[41] P. Florence, L. Manuelli, and R. Tedrake. Self-supervised correspondence in visuomotor policy learning. *IEEE Robotics and Automation Letters*, 5(2):492–499, 2019.

[42] M. Tsimpoukelli, J. L. Menick, S. Cabi, S. Eslami, O. Vinyals, and F. Hill. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212, 2021.

[43] K. Zakka, A. Zeng, P. Florence, J. Tompson, J. Bohg, and D. Dwibedi. Xirl: Cross-embodiment inverse reinforcement learning. In *Conference on Robot Learning*, pages 537–546. PMLR, 2022.

[44] J. Lu, D. Batra, D. Parikh, and S. Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

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

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

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

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

[49] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014.

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

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

[52] S. Karpagavalli and E. Chandra. A review on automatic speech recognition architecture and approaches. *International Journal of Signal Processing, Image Processing and Pattern Recognition*, 9(4):393–404, 2016.

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

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

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

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

[57] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

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

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

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

[61] Y. Su, T. Lan, Y. Liu, F. Liu, D. Yogatama, Y. Wang, L. Kong, and N. Collier. Language models can see: Plugging visual controls in text generation. *arXiv preprint arXiv:2205.02655*, 2022.

[62] Y. Tewel, Y. Shalev, I. Schwartz, and L. Wolf. Zero-shot image-to-text generation for visual-semantic arithmetic. *arXiv preprint arXiv:2111.14447*, 2021.

[63] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.

[64] R. Vedantam, C. Lawrence Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.

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

[66] A. Kamath, M. Singh, Y. LeCun, G. Synnaeve, I. Misra, and N. Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790, 2021.

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

[68] Speech-to-text: Automatic speech recognition | google cloud. `https://cloud.google.com/speech-to-text`. Accessed: 2022-05-13.

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

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

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

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

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

[74] P. Sharma, N. Ding, S. Goodman, and R. Soricut. 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)*, pages 2556–2565, 2018.

[75] Y. Li, T. Nagarajan, B. Xiong, and K. Grauman. 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*, pages 6943–6953, 2021.

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

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

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

[79] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. Davison, et al. 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*, pages 559–568, 2011.

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

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

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

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

[84] A. Zeng, P. Florence, J. Tompson, S. Welker, J. Chien, M. Attarian, T. Armstrong, I. Krasin, D. Duong, V. Sindhwani, et al. Transporter networks: Rearranging the visual world for robotic manipulation. *arXiv preprint arXiv:2010.14406*, 2020.

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

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

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

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

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

[90] J. Dong, X. Li, C. Xu, S. Ji, Y. He, G. Yang, and X. Wang. Dual encoding for zero-example video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9346–9355, 2019.

[91] J. Dong, X. Li, and C. G. Snoek. Predicting visual features from text for image and video caption retrieval. *IEEE Transactions on Multimedia*, 20(12):3377–3388, 2018.

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

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

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

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

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

[97] M. Bain, A. Nagrani, G. Varol, and A. Zisserman. 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*, pages 1728–1738, 2021.

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

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

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

[101] Y. J. Lee, J. Ghosh, and K. Grauman. Discovering important people and objects for egocentric video summarization. In *IEEE conference on computer vision and pattern recognition (CVPR)*, pages 1346–1353, 2012.

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

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

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

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

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

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

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

[109] G. Garcia-Hernando, S. Yuan, S. Baek, and T.-K. Kim. 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)*, pages 409–419, 2018.

[110] Y. Li, M. Liu, and J. M. Rehg. 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)*, pages 619–635, 2018.

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

[112] A. Furnari and G. M. Farinella. 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)*, pages 6252–6261, 2019.

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

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

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

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

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

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

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

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

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

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

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

[124] M. S. Ryoo. Human activity prediction: Early recognition of ongoing activities from streaming videos. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1036–1043, 2011.

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

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

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

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

[129] F. Abuzaid, G. Sethi, P. Bailis, and M. Zaharia. 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*, pages 1250–1261. IEEE, 2019. doi: 10.1109/ICDE.2019.00114. URL `https://doi.org/10.1109/ICDE.2019.00114`.

[130] A. Shrivastava and P. Li. Asymmetric LSH (ALSH) for sublinear time maximum inner product search (MIPS). In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2321–2329, 2014. URL `https://proceedings.neurips.cc/paper/2014/hash/310ce61c90f3a46e340ee8257bc70e93-Abstract.html`.

[131] K. M. Choromanski, V. Likhosherstov, D. Dohan, X. Song, A. Gane, T. Sarlós, P. Hawkins, J. Q. Davis, A. Mohiuddin, L. Kaiser, D. B. Belanger, L. J. Colwell, and A. Weller. Rethinking attention with performers. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL `https://openreview.net/forum?id=Ua6zuk0WRH`.

[132] H. Ramsauer, B. Schäfl, J. Lehner, P. Seidl, M. Widrich, L. Gruber, M. Holzleitner, T. Adler, D. P. Kreil, M. K. Kopp, G. Klambauer, J. Brandstetter, and S. Hochreiter. Hopfield networks is all you need. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL `https://openreview.net/forum?id=tL89RnzIiCd`.

[133] A. S. Rawat, J. Chen, F. X. Yu, A. T. Suresh, and S. Kumar. Sampled softmax with random fourier features. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13834–13844, 2019. URL `https://proceedings.neurips.cc/paper/2019/hash/e43739bba7cdb577e9e3e4e42447f5a5-Abstract.html`.

[134] K. Choromanski, H. Chen, H. Lin, Y. Ma, A. Sehanobish, D. Jain, M. S. Ryoo, J. Varley, A. Zeng, V. Likhosherstov, D. Kalashnikov, V. Sindhwani, and A. Weller. Hybrid random features. *to appear in ICLR 2022*, abs/2110.04367, 2021. URL `https://arxiv.org/abs/2110.04367`.

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

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

Appendix for *Socratic Models*

A Overview

The appendix includes: (i) unsupervised evaluation for model selection, (ii) additional notes on main experiments, (iii) more details on applications to egocentric perception, (iv) scaling video search for language-based world-state history, (v) more details on robot perception and planning experiments, (vi) additional discussion on future work (e.g., SMs for deductive reasoning) and (vii) broader impacts (e.g., energy and resource consumption). For code, see socraticmodels.github.io.

B 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?’ for SMs, we can instead ask: ‘(B): how well does this VLM compensate for the weakness of the LM?’.

Strope et al. [85] 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. [85] 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 [69]. 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

Table 5: 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. 5 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) [86] and ResNets (RN50) [87] 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) [3], as measured by zero-shot image classification accuracy on ImageNet [29] – 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 [85]), 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. 5.

C Additional Notes on Experiments

Choice of models. There are many options of large pretrained "foundation" [4] models to choose from, but our experiments in the main paper use models that are publicly available, so that our systems can be made accessible to the community. In particular, we use CLIP [3] as the text-image similarity VLM (ViT-L/14 with 428M params, except on MSR-VTT which uses ViT-B/32), ViLD [65] as the open-vocabulary object detector VLM; Wav2CLIP [88] as the sound-critic ALM and Google Cloud Speech-to-text API [68] as the speech-to-text ALM; GPT-3 with 175B params [2, 89] and RoBERTa [69] with 355M params as the LMs. All pretrained models are used off-the-shelf with no additional finetuning. In terms of compute resources required, all experiments can be run on a single machine using an NVIDIA V100 GPU with internet access for outsourced API calls (e.g., GPT-3 and Google Cloud Speech-to-text).

C.1 Image Captioning on MS COCO

For image captioning experiments on the MS COCO dataset [15, 16], we evaluate over a random sampled subset of 100 images from the test split [63], so that GPT-3 API runtime costs are more affordable for reproducibility (∼\$150 USD per run with with $n = 20$ generated candidate captions per image). Metrics (shown in Tab. 6) from baselines reported on this subset of MS COCO test examples are comparable to the full test set metrics. Also, while the captions in Fig. 3, Section 4.1, were generated with the prompt "…creative short…" as noted in Fig. 3, for best quantitative MS-COCO captions we used the prompt "…short, likely…".

Method	BLEU-4	METEOR	CIDEr	SPICE	ROUGE-L
*ClipCap [45] (full)	33.5	27.5	113.1	21.1	–
†MAGIC [61] (full)	12.9	17.4	49.3	11.3	39.9
ZeroCap [62] (full)	2.6	11.5	14.6	5.5	–
*ClipCap [45] (subset)	40.7	30.4	152.4	25.2	60.9
†MAGIC [61] (subset)	11.4	16.4	56.2	11.3	39.0
ZeroCap [62] (subset)	0.0	8.8	18.0	5.6	18.3

[*] finetuned on full training set with image-text pairs.
[†] finetuned on unpaired training set, zero-shot on image-text pairs.

Table 6: Image captioning metrics on the random subset of $N = 100$ (bottom) test examples are comparable to the full MS COCO test set metrics (top).

C.2 Contextual Image Captioning on Concadia

Our experiments on Concadia [17] evaluate the extent to which SMs can generate captions and descriptions conditioned on input images and their associated article text. While our results show that the SM combination of VLMs and LMs can achieve strong results on the benchmark, we also observe that LMs (e.g., GPT-3) alone can return surprisingly competitive results too (Tab. 7). Specifically, using the same LM prompt from the SM approach, but leaving out information from the VLM:

`I am an intelligent image captioning bot. The article is about: "{article_text}".` ~~`In this image, I think I see a {object1}, {object2}, {object3},...`~~ `A short caption for this image is:`

subsequently drops performance on image description by 2.0 CIDEr points, but surprisingly improves captioning performance by 1.2 CIDEr points. This suggests: (i) information from the VLM is more important for LMs in generating image descriptions than captions, (ii) there may be a strong correlation between the distributions of captions and article

Method	Caption Generation	Description Generation
Kreiss et al. [17]	11.3	17.4
SMs (ours)	38.9	22.6
SMs (no image)	40.1	20.6
SMs w/ description	93.8	–

Table 7: SMs on zero-shot contextual image captioning and description tasks on the Concadia dataset.

texts that can be leveraged by an LM alone, and/or (iii) there may exist overlap between Concadia (e.g., Wikipedia articles) and the training set of the LM, which warrants further investigation to disentangle confounding variables.

C.3 Video-to-text Retrieval on MSR-VTT 1k-A

We also report results in Tab. 8 on the popular MSR-VTT "1k-A" subset, introduced by Yu et al. [72] created via random sampling on the full test set. We follow the same evaluation protocol for video-to-text retrieval as used in prior work [55, 90, 91, 70], which reports the minimum rank among all valid text captions for a given video query, and each test video is associated with 20 captions.

		MSR-VTT 1k-A					
Category	Method	R@1↑	R@5↑	R@10↑	MdR↓	Audio	CLIP enc.
Finetuning	Collaborative Experts [55]	20.6	50.3	64.0	5.3	yes	
	SSB [92]	28.5	58.6	71.6	3.0	no	
	CLIP4Clip [93]	43.1	70.5	81.2	2.0	no	ViT-V/32
	CLIP2Video [71]	43.5	72.3	82.1	2.0	no	ViT-V/32
	DRL [94], ViT-B/32	45.3	73.9	83.3	2.0	no	ViT-V/32
	CAMoE [73]	49.1	74.3	84.3	2.0	no	ViT-B/32
	CLIP2TV [46]	54.1	77.4	85.7	**1.0**	no	ViT-B/16
	DRL [94], ViT-B/16 + QB-n	**56.2**	**79.9**	**87.4**	**1.0**	no	ViT-B/16
Zero-shot	SSB [92], zero-shot	8.7	23.0	31.1	31.0	no	
	CLIP via [67]	58.0	82.5	90.2	1.0	no	ViT-B/32
	SMs (ours)	**60.7**	**84.1**	**90.6**	**1.0**	yes	ViT-B/32

Table 8: Video-to-text retrieval results on MSR-VTT [18] dataset on the 1k-A [72] subset. 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 [3] is used, which CLIP encoder is used.

Note that the original CLIP baseline for video-to-text retrieval via Portillo-Quintero et al. [67] reports R@1 to be 27.2, but this was computed with only 1 caption per video that was random sampled [72] from the original set of 20 captions (for text-to-video retrieval). This differs from the original evaluation protocol and may be sub-optimal since the sampled caption can be ambiguous or partial (generated from crowd compute). For example, videos may be paired with a vague caption "a person is explaining something" as ground truth, rather than one of the other (more precise) captions e.g., "a person is talking about importing music to a ipod". Upon correcting the evaluation protocol (i.e., increasing the number of associated captions per video to 20), R@1 for Portillo-Quintero et al. [67] improves to 58.0, and SMs improve on top of that with LMs and ALMs[4] to 60.7 R@1 zero-shot.

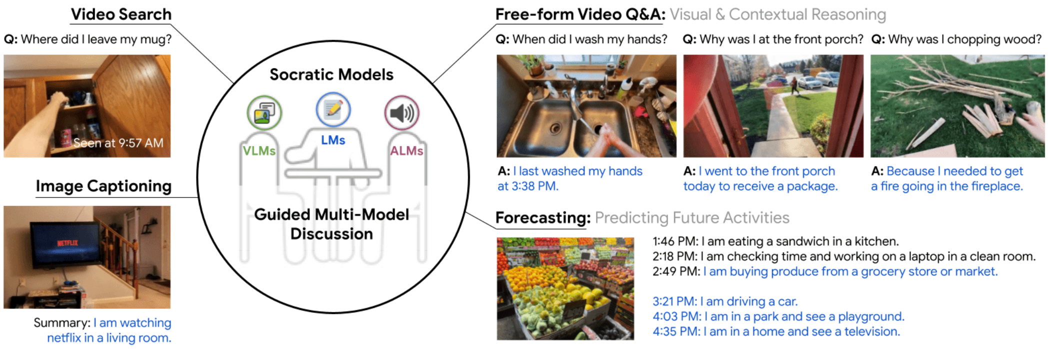
Other methods have also evaluated on zero-shot MSR-VTT *text-to-video* retrieval [95, 96, 97], but these have all been outperformed by Portillo-Quintero et al. [67]. Our method may be adapted as well to *text-to-video*, but due to our use of transcripts on only a subset of the videos, unlike in video-to-text, this creates an asymmetry which may require an unwieldy relative weighting for ranking videos with or without transcripts. Note that (Tab. 8) prior to the CLIP revolution in video-to-text retrieval, using the audio modality was not uncommon amongst competitive video-to-text retrieval methods [70, 55]. 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.

D Egocentric Perception Appendix

Background. 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) [98, 99] as well as features from human gaze, hands, and objects [100, 101, 102, 103, 104, 105]. 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 pretrained convolutional network features [106], finetuning

---

[4] Key used parameters for Google Cloud Speech-to-Text API 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.


Video Search
Image Captioning
Socratic Models
VLMs
LMs
ALMs
Guided Multi-Model Discussion
Free-form Video Q&A: Visual & Contextual Reasoning
Forecasting: Predicting Future Activities
Seen at 9:57 AM

Q: Where did I leave my mug?

Summary: I am watching netflix in a living room.

Q: When did I wash my hands? **A:** I last washed my hands at 3:38 PM.

Q: Why was I at the front porch? **A:** I went to the front porch today to receive a package.

Q: Why was I chopping wood? **A:** Because I needed to get a fire going in the fireplace.

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.

Figure 7: On various egocentric 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.

them [107, 48], or training them from scratch [108] 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 [109, 110]. [111] learns multimodal embeddings (i.e., video + audio), and [112] 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 [113], Charades-Ego [76], and Ego4D [19].

D.1 Why Egocentric Perception?

We highlight SMs on egocentric perception because it is an important yet challenging computer vision domain [19, 20, 76] with downstream applications in augmented reality (AR) and robotics [7]. 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 [29, 16, 74]. 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 [75, 76]. Overall, the key challenges 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 [76], Kinetics [114]), and reflect the information already stored in the models, 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.

D.2 Additional Details on 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. 8) 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. 8).

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

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.

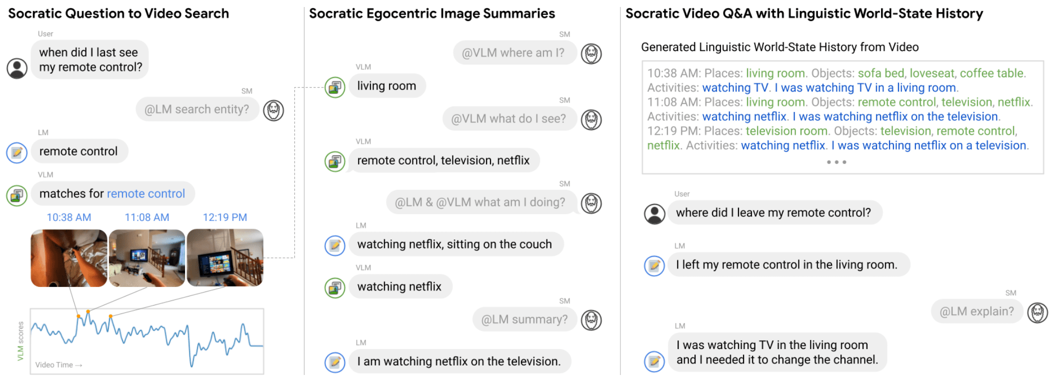
Temporal Reasoning Q&A

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 w 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.

Figure 8: 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 [45]. 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.


Socratic Question to Video Search
User
SM
LM
VLM
10:38 AM
11:08 AM
12:19 PM
VLM scores
Video Time →
Socratic Egocentric Image Summaries
Socratic Video Q&A with Linguistic World-State History
Generated Linguistic World-State History from Video

when did I last see my remote control?

@LM search entity?

remote control

matches for remote control

@VLM where am I?

living room

@VLM what do I see?

remote control, television, netflix

@LM & @VLM what am I doing?

watching netflix, sitting on the couch

watching netflix

@LM summary?

I am watching netflix on the television.

10:38 AM: Places: living room. Objects: sofa bed, loveseat, coffee table. Activities: watching TV. I was watching TV in a living room.
11:08 AM: Places: living room. Objects: remote control, television, netflix. Activities: watching netflix. I was watching netflix on the television.
12:19 PM: Places: television room. Objects: television, remote control, netflix. Activities: watching netflix. I was watching netflix on a television.
• • •

where did I leave my remote control?

I left my remote control in the living room.

@LM explain?

I was watching TV in the living room and I needed it to change the channel.

Figure 9: 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.

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 rank Places365 [59] 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 [115] 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 [116] 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 [116]. 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. 8, and compare them to results from a state-of-the-art image captioning model, ClipCap [45]. 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 [117], which aligns with our observations. Relatively less research has been carried out specifically on egocentric image captioning [118]. 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 [88] as the ALM. Wav2CLIP is trained on 5-second audio clips from the VGGSound dataset [119], 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. 10 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. 10 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:

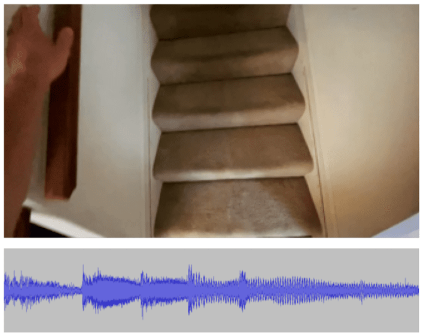


Figure 10: Example frame and corresponding (centered) 5-second audio clip which provide the driving example for Sec. D.2-B, i.e., adding in ALMs into Socratic dialogue to improve single-moment summarization. Note that this waveform mostly represents the background piano music, but the system is still 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 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. D.3) 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.

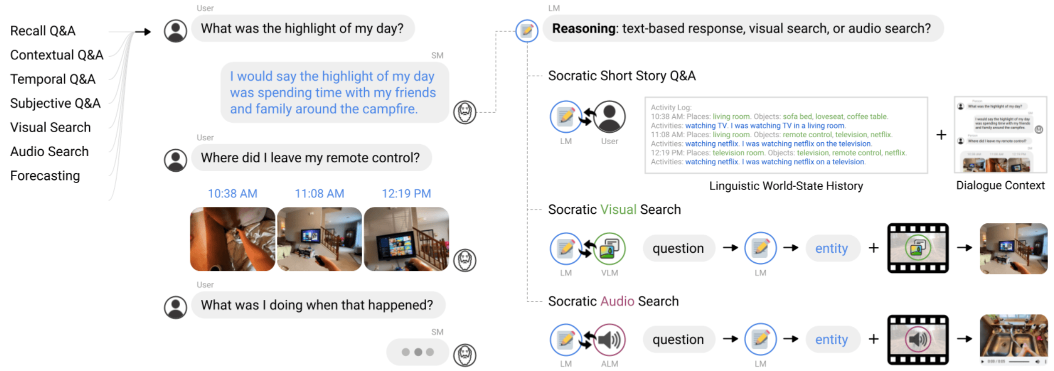
3.3-C. Compiling a Language-Based World-State History

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 [78], 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 ([120]) 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. 8, middle, shows results generated by our system. The specific event log shown in Fig. 8 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.



What was the highlight of my day?

I would say the highlight of my day was spending time with my friends and family around the campfire.

Where did I leave my remote control?

What was I doing when that happened?

Activity Log:
10:38 AM: Places: living room. Objects: sofa bed, loveseat, coffee table.
Activities: watching TV. I was watching TV in a living room.
11:08 AM: Places: living room. Objects: remote control, television, netflix.
Activities: watching netflix. I was watching netflix on the television.
12:19 PM: Places: television room. Objects: television, remote control, netflix.
Activities: watching netflix. I was watching netflix on a television.

Figure 11: 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.

D.3 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. 7). There are of course limitations to what they can provide, but our demonstrated examples suggest that we can already today 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. D.2), 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.

(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). 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 [121] for an early survey and [122, 123] 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 [81] we can further 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 qualitative examples of free-form question answering using our SM system on egocentric video in Fig. 8, bottom, Fig. 9, and Fig. 11 generated using a first-person POV video[5] 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 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 [10]. 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 predictions can be preemptively specified depending on application needs. The completion results are generative, and more broad than binary event classification (e.g., [82]). Example completion (also shown in Fig. 7):

---

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

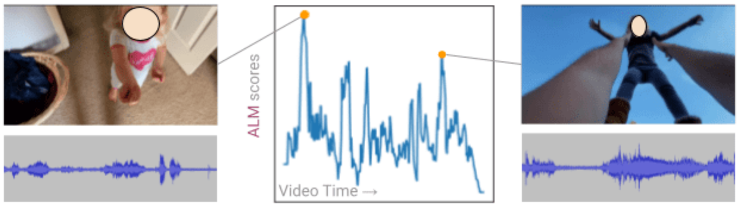


Figure 12: Example zero-shot language-prompted auditory retrieval (shown: top 2 results) in response to "what did my daughter's laugh sound like today?", for which an LM identifies the audio search query of "daughter's laugh", and an ALM (Wav2CLIP) is used for audio retrieval. The top (left) retrieval 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.

```
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 [124, 125, 126], human locations [127], or image features [128] 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:

```
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 ([120]) using ALM features for "daughter's laugh" (Fig. 12). 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 [65]. 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.

E Scaling Up Socratic Video Search

The search algorithms of the SMs, which may be used both for compiling world-state history (Sec. D.2-C) and for video search retrieval (Sec. D.3) 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 pre-processing 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: [129]). Those include pruning and various indexing techniques, such as LSH-hashing [130]. 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 [131] combined with *modern continuous associative memory* (MCAM) models [132]. 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* [131]. 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 [133]. This approach relaxes the MIP-search to sampling from the linearized softmax distribution via FAVOR+ [134]. 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.

F Additional Notes on Robot Experiments

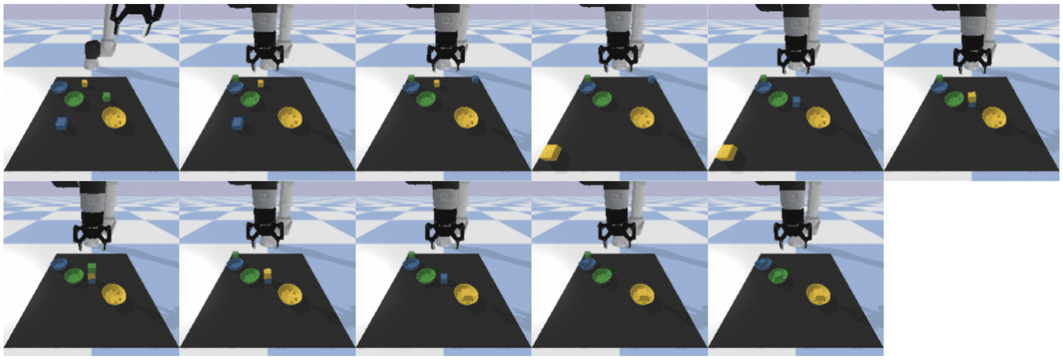
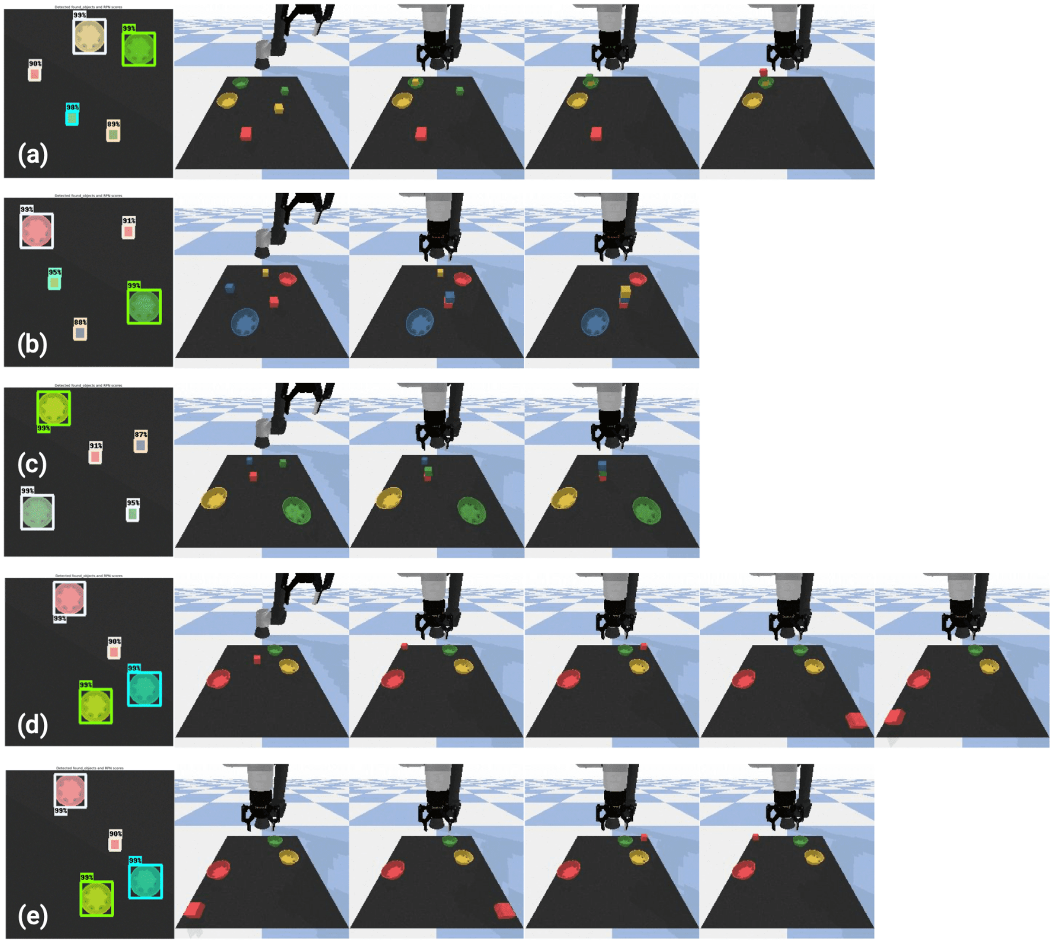


Figure 13: Full rollout of the robot environment for the example presented in Sec. 5.3 of the main paper.

The SM robot system uses a VLM (open-vocabulary object detection with ViLD [65]) to describe the objects in the scene, feeds that description as context to a LM as a multi-step planner [7, 6], that then takes as input a natural language instruction and generates the individual steps to be passed to a pretrained language-conditioned robot policy, for which we specifically use a CLIP-conditioned [83] No-Transport baseline from Zeng et al. [84] (inspired by CLIPort [83] for open vocabulary pick-and-place). The full prompt used as context to the LM for multi-step planning is:

```
objects = [”cyan block”, ”yellow block”, ”brown block”, ”green bowl”]
# move all the blocks to the top left corner.
Step 1. robot.pick_and_place(”brown block”, ”top left corner”)
Step 2. robot.pick_and_place(”cyan block”, ”top left corner”)
Step 3. robot.pick_and_place(”yellow block”, ”top left corner”)
# put the yellow one the green thing.
Step 1. robot.pick_and_place(”yellow block”, ”green bowl”)
# undo that.
Step 1. robot.pick_and_place(”yellow block”, ”top left corner”)
objects = [”pink block”, ”gray block”, ”orange block”]
# move the pinkish colored block on the bottom side.
Step 1. robot.pick_and_place(”pink block”, ”bottom side”)
objects = [”orange block”, ”purple bowl”, ”cyan block”, ”brown bowl”, ”pink block”]
# stack the blocks.
Step 1. robot.pick_and_place(”pink block”, ”orange block”)
Step 2. robot.pick_and_place(”cyan block”, ”pink block”)
# unstack that.
Step 1. robot.pick_and_place(”cyan block”, ”bottom left”)
Step 2. robot.pick_and_place(”pink block”, ”left side”)
objects = [”red block”, ”brown block”, ”purple bowl”, ”gray bowl”, ”brown bowl”, ”pink block”, ”purple block”]
# group the brown objects together.
Step 1. robot.pick_and_place(”brown block”, ”brown bowl”)
objects = [”orange bowl”, ”red block”, ”orange block”, ”red bowl”, ”purple bowl”, ”purple block”]
# sort all the blocks into their matching color bowls.
Step 1. robot.pick_and_place(”orange block”, ”orange bowl”)
Step 2. robot.pick_and_place(”red block”, ”red bowl”)
Step 3. robot.pick_and_place(”purple block”, ”purple bowl”)
```

Fig. 13 depicts a full rollout of the example in Sec. 5.3 in the main paper, which involves human dialogue. Fig. 14 shows additional examples of multi-step tasks that the system can perform out-of-the-box with zero-shot SMs. The system is able to reason over order and nuanced language (clockwise vs. counterclockwise) as well as respond to different objects being detected (the stacking task with varied block colors). Note that the LM is few-shot prompted to generate pseudo code “robot.pick_and_place("A", "B")” which calls a function to return a fixed template sentence “Pick the A and place it on the B.” subsequently fed as input to the language-conditioned robot policy. While we can prompt the LM to directly produce the template sentences as opposed to code, we find that the LM can sometimes generate phrases or prepositions that are beyond the training set of the language-conditioned policy. We observe that the policies are more likely to return correct actions when the templates can be engineered to be more similar to the phrases seen within the policy's training data. We also found ViLD and CLIP to be brittle in this scene, as the scene is simulated and the objects are not natural. High-performance in this setting requires a good view angle (an overhead camera), filtered colors (red, green, yellow, and blue), and tuned names (we referred to the blocks as “boxes” and the bowls as “circles” to account for the overhead view). Without these changes, the system still is able to complete many tasks, but less consistently. We expect off-the-shelf VLMs of the future to be more robust than those currently available.



```
(a)
objects = [’green bowl’, ’yellow bowl’, ’yellow block’, ’green block’, ’red block’]
# Put all the blocks in the green bowl.
Step 1. robot.pick_and_place(’yellow block’, ’green bowl’)
Step 2. robot.pick_and_place(’green block’, ’green bowl’)
Step 3. robot.pick_and_place(’red block’, ’green bowl’)

(b)
objects = [’red bowl’, ’blue bowl’, ’yellow block’, ’red block’, ’blue block’]
# Stack all the blocks.
Step 1. robot.pick_and_place(’blue block’, ’red block’)
Step 2. robot.pick_and_place(’yellow block’, ’blue block’)

(c)
objects = [’green bowl’, ’yellow bowl’, ’green block’, ’red block’, ’blue block’]
# Stack all the blocks.
Step 1. robot.pick_and_place(’green block’, ’red block’)
Step 2. robot.pick_and_place(’blue block’, ’green block’)

(d)
objects = [’green bowl’, ’red bowl’, ’yellow bowl’, ’red block’]
# Clockwise, move the block through all the corners.
Step 1. robot.pick_and_place(’red block’, ’top left corner’)
Step 2. robot.pick_and_place(’red block’, ’top right corner’)
Step 3. robot.pick_and_place(’red block’, ’bottom right corner’)
Step 4. robot.pick_and_place(’red block’, ’bottom left corner’)

(e)
# Now, move it back counterclockwise.
Step 5. robot.pick_and_place(’red block’, ’bottom left corner’)
Step 6. robot.pick_and_place(’red block’, ’bottom right corner’)
Step 7. robot.pick_and_place(’red block’, ’top right corner’)
Step 8. robot.pick_and_place(’red block’, ’top left corner’)
```

Figure 14: Additional examples of multi-step tasks that the SM robot system can perform out-of-the-box.

G Socratic Deductive Reasoning

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., the dog went into the kitchen, the groceries are still in the kitchen, and 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 [56] 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.

H Broader Impact: Energy and Resource Consumption

Regarding the impact on energy and other resource consumption, 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 circuitry [135], optical diffraction [136]) for inference with substantially lower power consumption and more compact form factors.