

Socratic Models: Composing Zero-Shot Multimodal Reasoning with Language

Robotics and AR at Google

Выполнил Карлов В.А.
БПМИ193

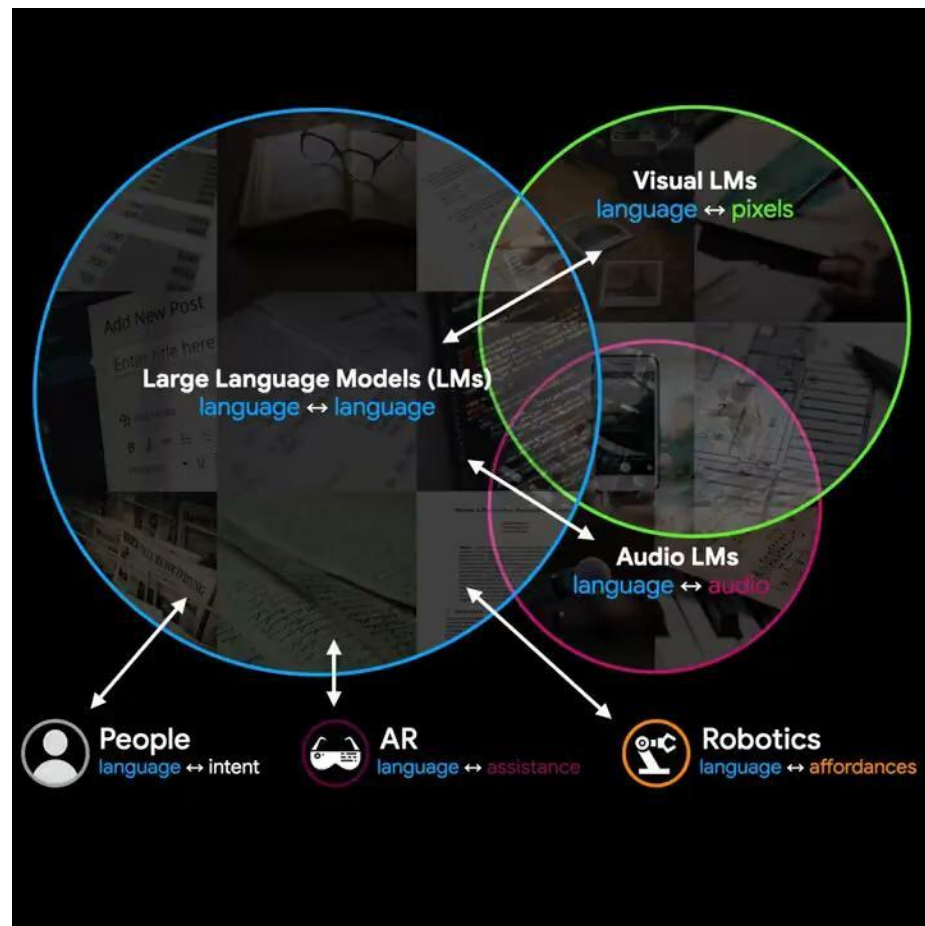
Постановка задачи

Создать следующий фреймворк:

- несколько предобученных моделей
- данные разной модальности
- без дополнительного обучения (zero-shot composed)
- выход – единый текст-описание

Цвета:

- **ТЕКСТ**
- **ВИЗУАЛЬНАЯ ИНФОРМАЦИЯ**
- **ЗВУК**



1.1. Image Captioning

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.

$$\text{caption} = f_{\text{VLM}}^3(f_{\text{LM}}^2(f_{\text{VLM}}^1(\text{image})))$$

1. **VLM** object detection: place, image type, number of people...
2. Top-k results \rightarrow **LM** prompt
3. **LM** for text completion (+ noise for diverse) \rightarrow n candidate captions
4. **VLM** ranking \rightarrow highest scoring caption

1.1. Image Captioning – результаты



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.

1.1. Image Captioning – результаты на MS COCO

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.

1.2. Video-to-Text Retrieval – демонстрация

Zero-Shot
Socratic
Video-to-Text
Retrieval

1.2. Video-to-Text Retrieval

- Берем готовый VTR-фреймворк 2021 года (Portillo-Quintero et al):
 - VLM features of all video frames per video → average
 - Text features of captions
 - One-to-many NN matching
- Добавляем к нему:
 - ASR (e.g. Google Cloud speech-to-text API)
 - Суммаризируем полученный транскрипт с помощью GPT-3
- Итоговый **matching score** =
$$\text{CLIP}(\text{caption}) \cdot \text{CLIP}(\text{video}) \times$$
$$\text{RoBERTa}(\text{caption}) \cdot \text{RoBERTa}(\text{GPT-3}(\text{prompt}, \text{Speech2Text}(\text{audio})))$$

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,

1.2. Video-to-Text Retrieval – результаты

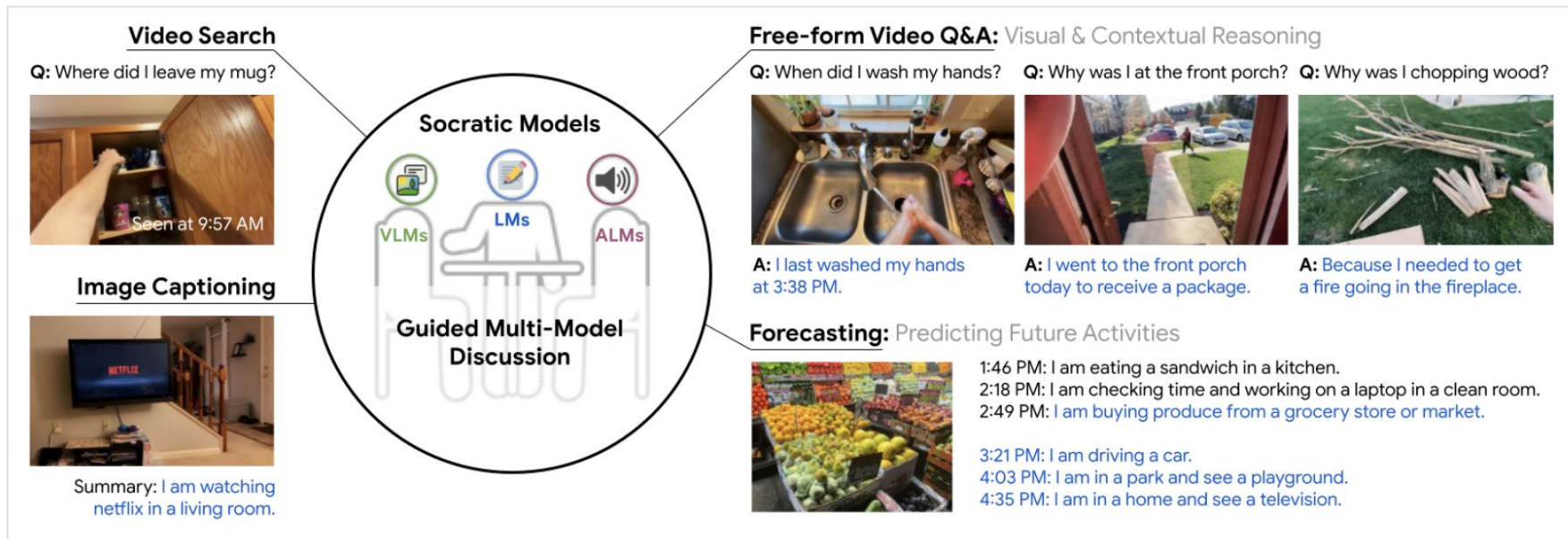
Category	Method	MSR-VTT Full				Audio
		R@1↑	R@5↑	R@10↑	MdR↓	
<i>Finetuned</i>	JEMC [70]	12.5	32.1	42.4	16.0	yes
	Collab. Experts [55]	15.6	40.9	55.2	8.3	yes
	CLIP2Video [71]	54.6	82.1	90.8	1.0	no
<i>Zero-shot</i>	CLIP via [67]	40.3	69.7	79.2	2.0	no
	SMs (ours)	44.7	71.2	80.0	2.0	yes

– comparing on whole MSR-VTT dataset (new Zero-Shot SOTA)

<i>Long-transcript subset of</i> 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

– comparing on long-transcript subset (substantial increase)

2.1. Egocentric perception



Задачи рамок восприятия от первого лица:

- суммаризация контента / поиск по видео
- ответы на вопросы в свободной форме
- предсказание будущих действий

2.1. Egocentric perception

Как это работает?

- идея: видео → текст
- ключевые моменты / фреймы (например, по исходному запросу)
- их текстовое описание (аналогично 1-му разделу)
- рекурсивно суммаризируем
- получаем текстовое описание событий ролика



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.**

2.1. Egocentric perception – демонстрация



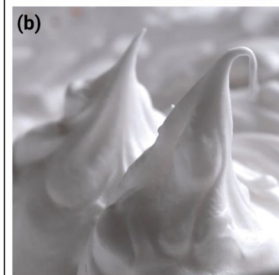
8:31 AM: Places: clean room. Objects:
shorts, jeans, shirt. Commonsense
activities: getting dressed. Most likely:
getting dressed. I was getting dressed,

Language-based World-state History

2.2. Multimodal Assistive Dialogue

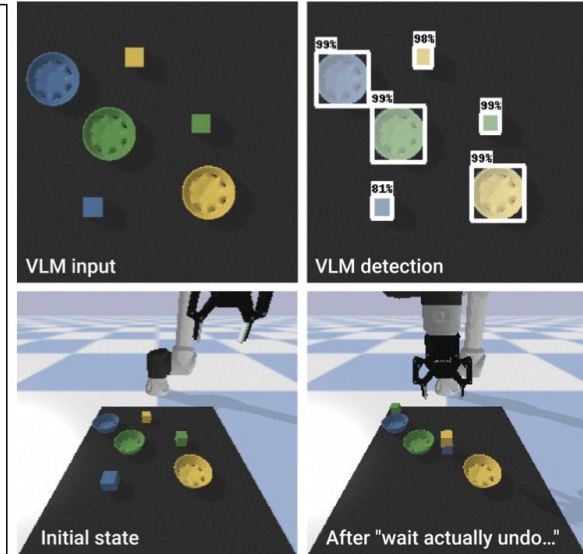
- GPT-3
- Query for visual information from user
→ web crawler
- VLM runs visual search → retrieve a related frame

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.



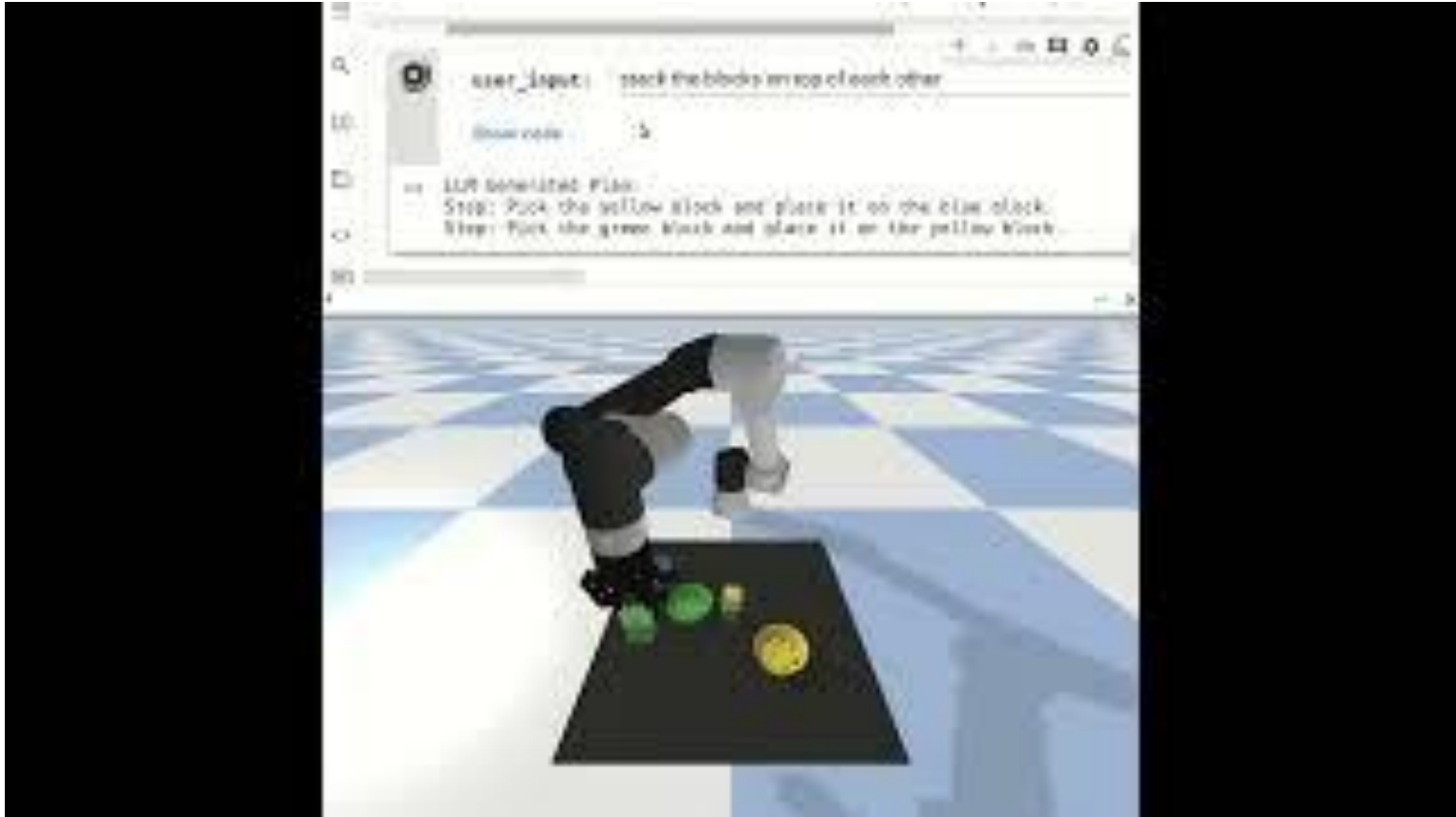
2.3. Robot Perception & Planning

```
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")
```



1. **VLM** (object detection) → description of the objects in the scene
2. Description as context → **LM** as a multi-step planner
3. Individual steps → pretrained language-conditioned **robot policy**

2.3. Robot Perception & Planning – демонстрация



Спасибо за внимание!