

How Does Cognitive Ability Influence Visual Understanding? How Do We Construct Cognitive Abilities in Vision-Language Models?



Jiapeng Li
2023/06/15

How Does Cognitive Ability Influence Visual Understanding?

(LLAVA)VQA:

What is unusual about this image?



The unusual aspect of this image is that President Obama is standing on a scale and weighing himself while being watched by a group of men, including a man in a mirror. This is an unconventional scene, as one would typically expect to see a President engaged in more formal or official activities rather than stepping on a scale in a public setting. The presence of multiple men in the bathroom also adds to the peculiar nature of the scene.

(BLIP2)Caption:



Computation time on Intel Xeon 3rd Gen Scalable cpu: cached

a man in a suit and tie standing in front of a group of people

Combining the Visual Contexts Obtained

Understanding a Character's Intent

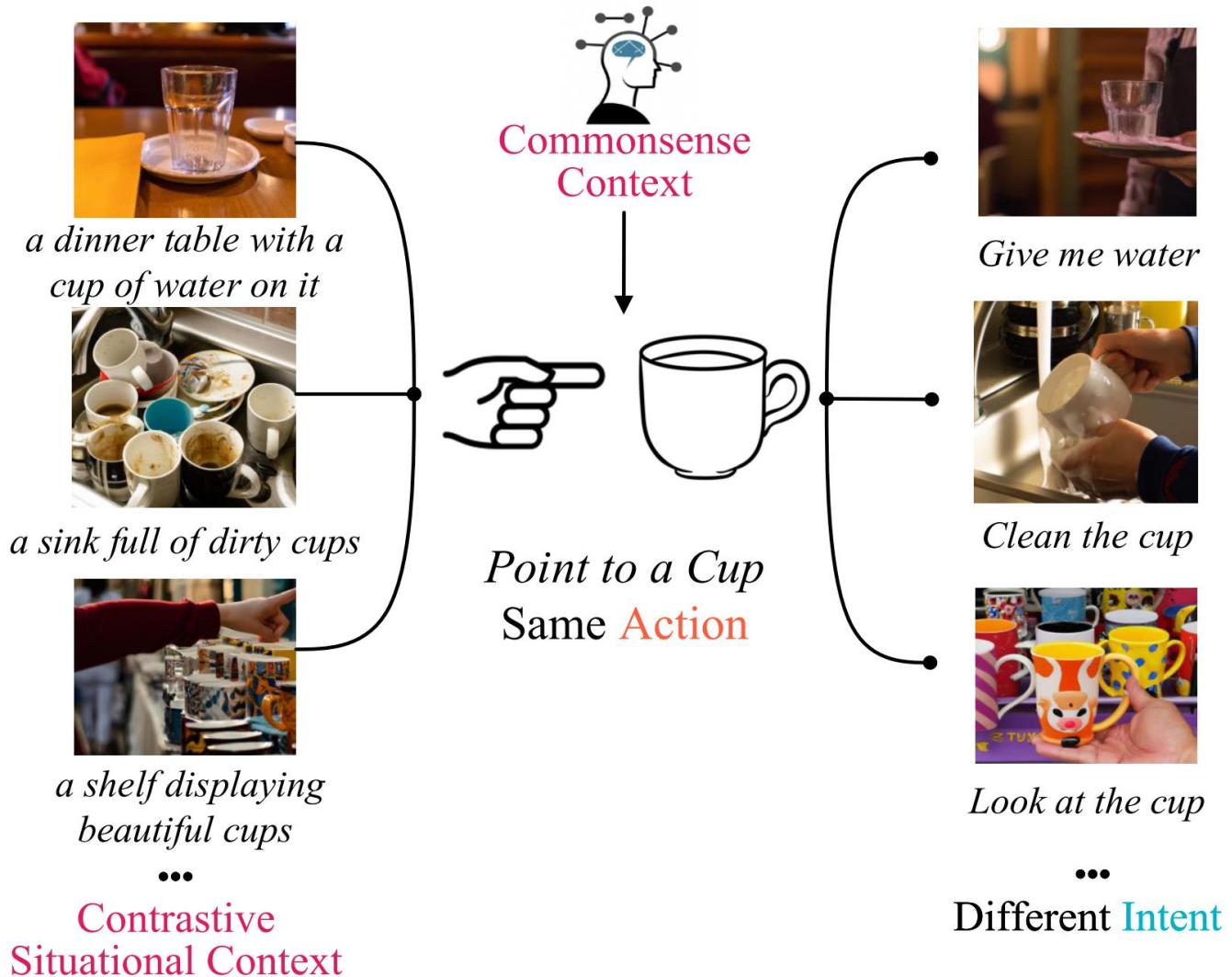
Understanding visual contexts from the perspective of intent is a unique cognitive ability of humans.

Model: 一个人站在一群人面前/一个在照镜子

Human: 奥巴马在偷偷踩前面人的体重秤

Our Work: IntentQA Context-aware Video Intent Reasoning

How Does Cognitive Ability Influence Visual Understanding?



The way humans understand intent is by making high-level abstractions of visual contexts from various aspects

IntentQA Dataset: QA that Infers Intent in 4 Different Ways.

Causal Why
(CW)



Q&A: Why did the man **point** to the screen when talking to the child?

- 0. looking at the path
- 1. draw child's attention**
- 2. check the hair cut
- 3. asking something
- 4. talking

Causal How
(CH)



Q&A: How did the man in light blue jacket show that he **wanted a drink**?

- 0. point his finger**
- 1. puts down his phone
- 2. bend down
- 3. rest on leg
- 4. with his hands

Temporal Next
(TN)



Q&A: What did the baby do after the boy **pointed** near the middle of the video?

- 0. shake hands
- 1. continue eating snacks
- 2. look at what he is pointing**
- 3. closes eyes
- 4. point in same direction

Temporal Previous
(TP)



Q&A: what did the baby do at the start before the girl in white **pointed**?

- 0. point to green spade
- 1. supervise
- 2. spit**
- 3. say something to boy
- 4. feed horse with grass

Model

Negative Sample

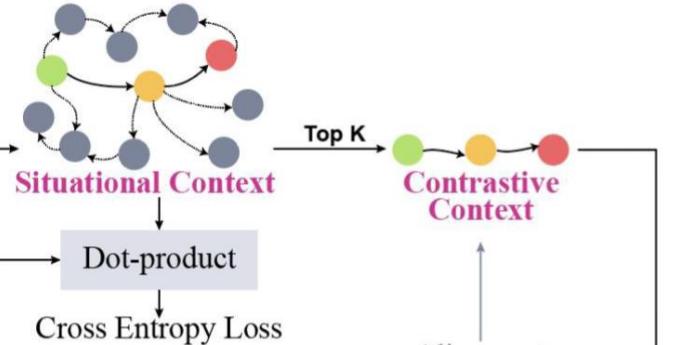
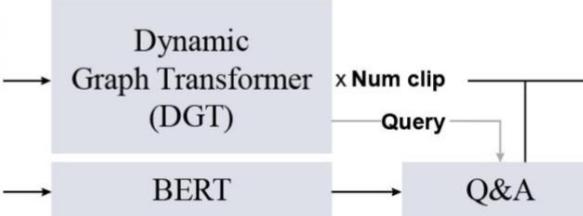
Q&A: Why does the lady in green use her finger to *point* to something?



Intent: Distract children

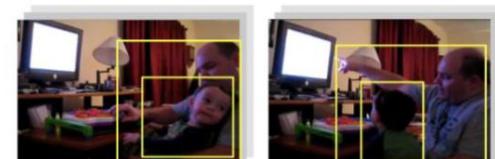


Question + Multiple choice: Answer $\times 5 \rightarrow$



Anchor Sample

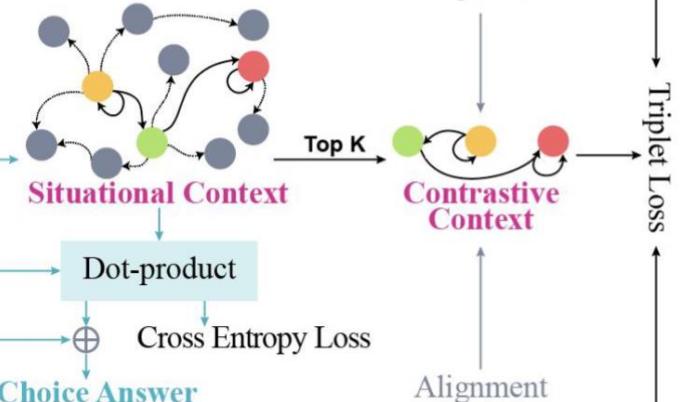
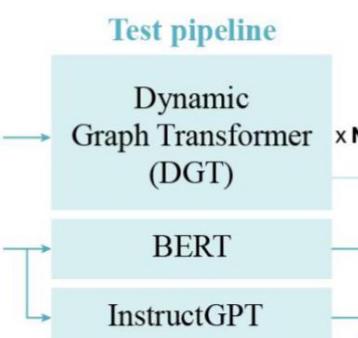
Q&A: Why did the man *point* to the screen when talking to the child?



Intent: Draw child's attention

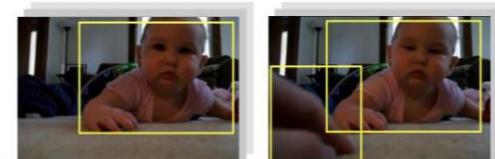


Question + Multiple choice: Answer $\times 5 \rightarrow$



Positive Sample

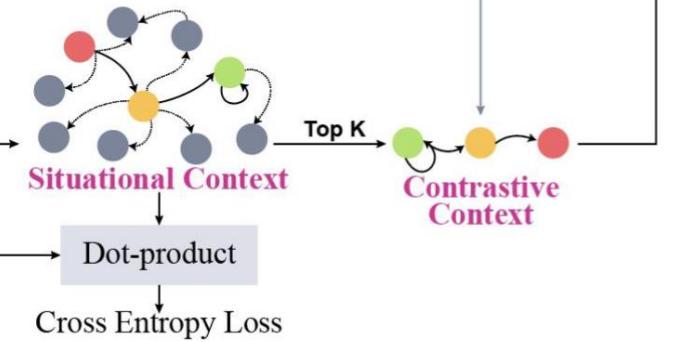
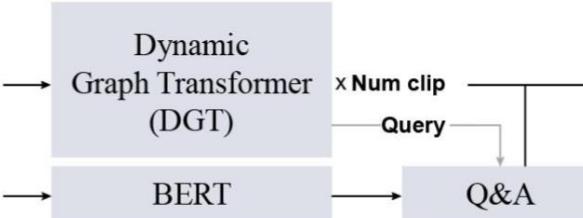
Q&A: Why does the person *point* to a place to the baby before the baby keeps coming closer and closer to the camera?



Intent: Catching baby's attention



Question + Multiple choice: Answer $\times 5 \rightarrow$



Experiments

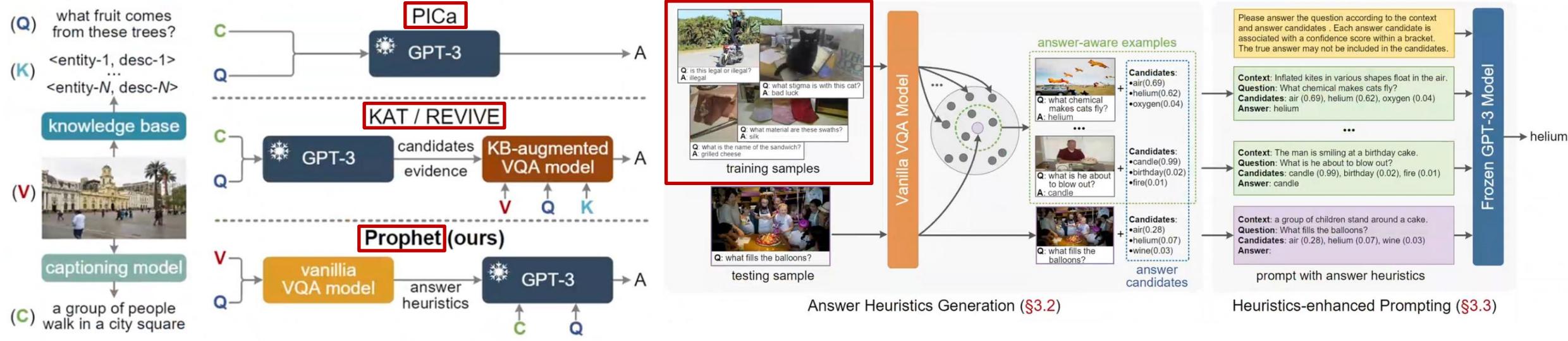
Model ID	Model	Text Rep.	CW		CH		TP&TN		Total	
			Val.	Test	Val.	Test	Val.	Test	Val.	Test
-	EVQA	GloVe	25.99	25.92	37.43	34.54	28.00	25.52	28.38	27.27
-	CoMem	GloVe	31.56	30.00	35.63	28.69	28.57	28.95	31.46	29.52
-	HGA	GloVe	29.45	32.00	35.03	30.64	29.71	31.05	30.43	31.54
-	HME	GloVe	30.97	34.40	35.33	34.26	34.29	29.14	32.53	33.08
-	HQGA	GloVe	32.49	33.20	38.32	34.26	34.48	36.57	33.95	34.21
-	CoMem	BERT	46.75	47.68	57.49	54.87	41.71	39.05	47.21	46.77
-	HGA	BERT	43.54	44.88	56.89	50.97	42.48	39.62	45.45	44.61
-	HME	BERT	46.50	46.08	51.20	54.32	44.76	40.76	46.82	46.16
-	HQGA	BERT	45.91	48.24	57.78	54.32	44.76	41.71	47.55	47.66
-	VGT	BERT	50.46	51.44	55.99	55.99	48.19	47.62	50.78	51.27
-	InstructGPT	-	-	52.16	-	61.28	-	43.43	-	51.55
-	GPT4	-	-	55.20	-	63.23	-	47.24	-	54.59
4	Ours w/o GPT	BERT	52.57	<u>55.28</u>	57.47	<u>61.56</u>	46.10	<u>47.81</u>	51.71	<u>54.50</u>
5	Ours + instructGPT	BERT	-	58.40	-	65.46	-	50.48	-	57.64
5	Ours + GPT4	BERT	-	60.00	-	65.74	-	53.14	-	59.28
-	Human	-	-	77.76	-	80.22	-	79.05	-	78.49

1. GPT and LLM models exhibit robust reasoning skills and can achieve results on par with leading models using only question contexts and inherent commonsense.
2. Even basic integration of our model with GPT for commonsense provision improves performance significantly, indicating the promising potential of LLM for enhancing V-L task outcomes.

How to Utilize LLM in V-L Models

How to Utilize LLM in V-L Models

in-context learning



PICa

KAT/REVIVE

Prophet

Liangke Gui, Borui Wang, Qiuyuan Huang, Alex Hauptmann, Yonatan Bisk, and Jianfeng Gao. Kat: A knowledge augmented transformer for vision-and-language. NAACL, 2021.

hengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. An empirical study of gpt-3 for few-shot knowledge-based vqa. In AAAI, pages 3081–3089, 2022

uanze Lin, Yujia Xie, Dongdong Chen, Yichong Xu, Chenguang Zhu, and Lu Yuan. REVIVE: Regional visual representation matters in knowledge-based visual question answering. In NeurIPS, 2022

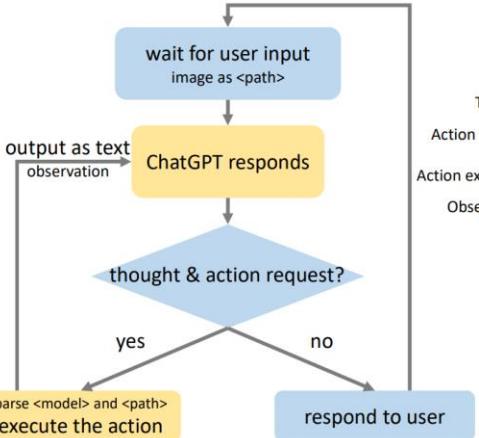
1. All models use in-context learning when prompting GPT.

2. KAT/REVIVE uses a VQA model they trained with a Knowledge Base as the decision model, while PICa and Prophet use GPT as the decision model.

3. Prophet, when choosing in-context learning examples, adopts a strategy similar to mine, where it finds more relevant samples based on candidate answers as examples for in-context learning.

How to Utilize LLM in V-L Models

- image caption
- image tagging
- dense caption
- celebrity
- receipt
- OCR
- Bing search
- ⋮
- vision experts



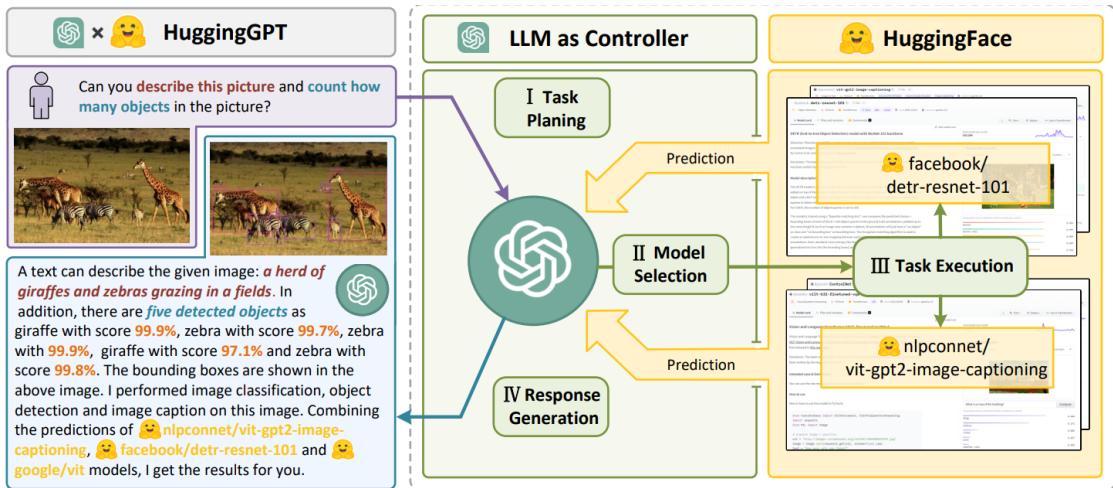
MM-REACT

Yang Z, Li L, Wang J, et al. Mm-react: Prompting chatgpt for multimodal reasoning and action[J]. arXiv preprint arXiv:2303.11381, 2023.

Shen Y, Song K, Tan X, et al. Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface[J]. arXiv preprint arXiv:2303.17580, 2023.

Surís D, Menon S, Vondrick C. Vipergpt: Visual inference via python execution for reasoning[J]. arXiv preprint arXiv:2303.08128, 2023.

Wu C, Yin S, Qi W, et al. Visual chatgpt: Talking, drawing and editing with visual foundation models[J]. arXiv preprint arXiv:2303.04671, 2023.



HuggingGPT

Query: What did the boy do after he dropped the sparkles on the floor?

Generated code

```

def execute_command(video, question, possible_answers):
    video_segment = VideoSegment(video)
    drop_detected = False
    for i, frame in enumerate(video_segment.frame_iterator()):
        if frame.exists("boy") and frame.exists("sparkles") and \
            frame.simple_query("is the boy dropping the sparkles?") == "yes":
            drop_detected = True
            break
    if drop_detected:
        index_frame = i + 1
    else:
        index_frame = video_segment.num_frames // 2
    frame_of_interest = ImagePatch(video_segment, index_frame)
    boys = frame_of_interest.find("boy")
    if len(boys) == 0:
        boys = [frame_of_interest]
    boy = boys[0]
    caption = boy.simple_query("What is this?")
    info = {
        "Caption of frame after dropping the sparkles": caption,
    }
    answer = select_answer(info, question, possible_answers)
    return answer
  
```

Execution



```

In: [Three frames showing a boy dropping sparkles]

frame.exists("boy") and \
frame.exists("sparkles") and \
frame.simple_query("is the boy dropping the sparkles?") == "yes":
  ► frame = {ImagePatch}

boys = frame_of_interest.find("boy")
► boy = {ImagePatch}

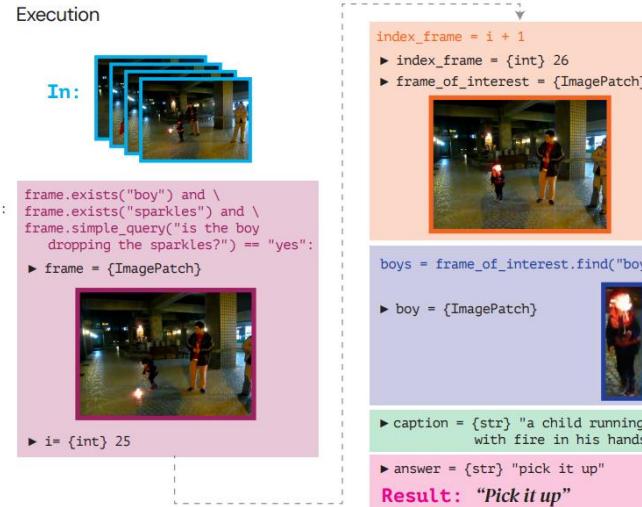
caption = {str} "a child running with fire in his hands"
► answer = {str} "pick it up"
Result: "Pick it up"
  
```

ViperGPT

Query: What did the boy do after he dropped the sparkles on the floor?

Generated code

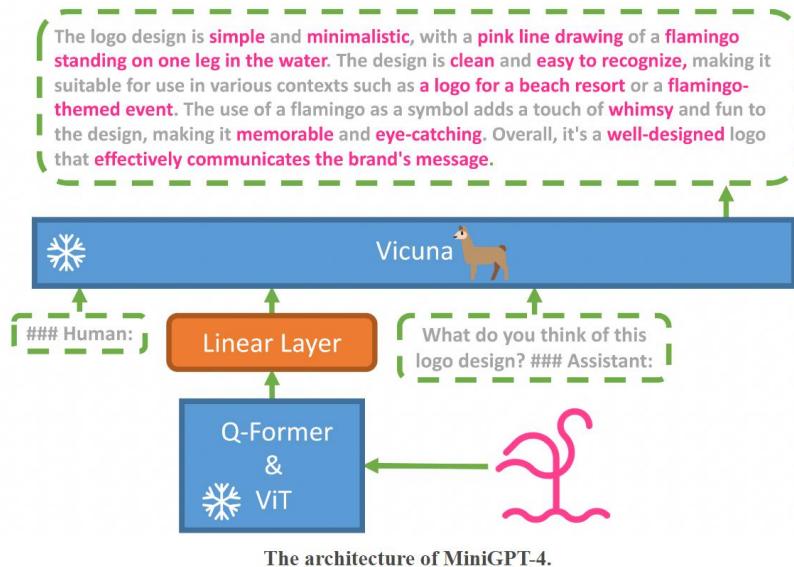
```
def execute_command(video, question, possible_answers):
    video_segment = VideoSegment(video)
    drop_detected = False
    for i, frame in enumerate(video_segment.frame_iterator()):
        if frame.exists("boy") and frame.exists("sparkles") and \
            frame.simple_query("is the boy dropping the sparkles?") == "yes":
            drop_detected = True
            break
    if drop_detected:
        index_frame = i + 1
    else:
        index_frame = video_segment.num_frames // 2
    frame_of_interest = ImagePatch(video_segment, index_frame)
    boys = frame_of_interest.find("boy")
    if len(boys) == 0:
        boys = [frame_of_interest]
    boy = boys[0]
    caption = boy.simple_query("What is this?")
    info = {
        "Caption of frame after dropping the sparkles": caption,
    }
    answer = select_answer(info, question, possible_answers)
return answer
```



1. Like earlier research, GPT is still the decision model.
2. Although video-capable, all vision models are really image models, using question-based preprocessing code to handle video.
3. Codex's closure by OpenAI has made it less reproducible, often generating nonfunctional code.
4. Code generated by GPT, which hasn't viewed the video, can miss context present in the video but not in the question. A recursive correction mechanism is needed.

1. **find(GLIP[31])**: This function takes an image and a short noun phrase as input and returns a list of image patches containing the specified object. This function is from the GLIP model.
2. **exists(GLIP[31])**: This function takes an image and a short noun phrase as input and returns a boolean value indicating whether the specified object is present in the image. This function is also from the GLIP model.
3. **verify_property(X-VLM[64])**: This function takes an image, a noun phrase representing an object, and an attribute representing a property of that object as input. It returns a boolean value indicating whether the property is present in the image. This function is from the X-VLM model.
4. **best_image_match(X-VLM[64])**: This function takes a list of image patches and a short noun phrase as input and returns the image patch that best matches the specified object. This function is also from the X-VLM model.
5. **best_text_match**: This function takes a list of noun phrases and one image as input and returns the noun phrase that best matches the image. This module is not necessary for visual grounding but is included for tasks with text outputs. It is implemented using an image-text similarity model similar to CLIP[43].
6. **compute_depth(MiDaS[44])**: This function computes the median depth of the image patch using the MiDaS model.
7. **distance**: This function computes the pixel-distance between two image patches using built-in Python tools.
8. **select_answer(GPT-3[6])**: given textual information about a scene and a list of possible answers, returns the answer that best fits the information. This function is associated with the GPT-3 model, which is designed for various natural language processing tasks, including question-answering and text completion.
9. **simple_query(BLIP-2[31])** handles basic queries that are not further decomposable, such as "What animal is this?"

How to Utilize LLM in V-L Models



The architecture of MiniGPT-4.

MiniGPT4

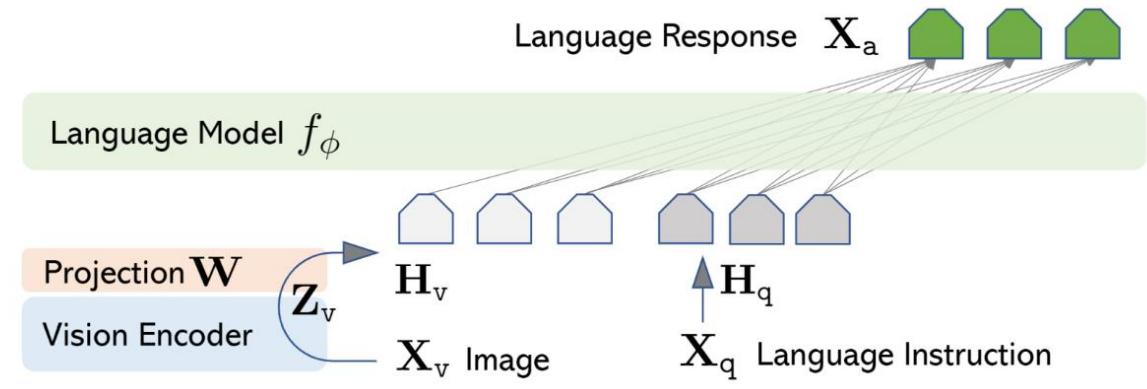
Zhu D, Chen J, Shen X, et al. Minigpt-4: Enhancing vision-language understanding with advanced large language models[J]. arXiv preprint arXiv:2304.10592, 2023.

Li K C, He Y, Wang Y, et al. Videochat: Chat-centric video understanding[J]. arXiv preprint arXiv:2305.06355, 2023.

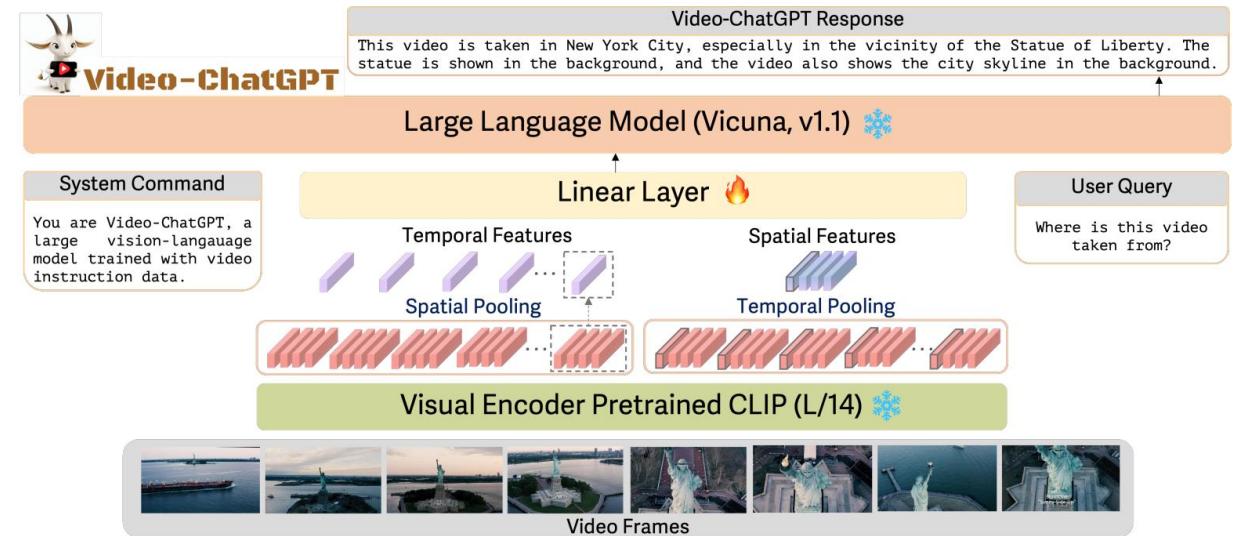
Liu H, Li C, Wu Q, et al. Visual instruction tuning[J]. arXiv preprint arXiv:2304.08485, 2023.

Maaz M, Rasheed H, Khan S, et al. Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models[J]. arXiv preprint arXiv:2306.05424, 2023.

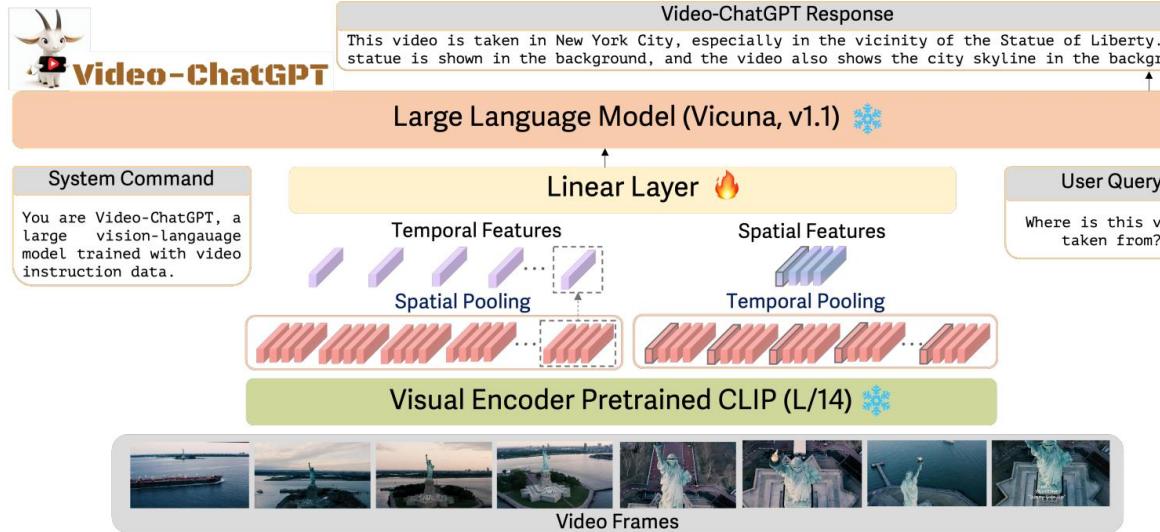
Dai W, Li J, Li D, et al. Instructblip: Towards general-purpose vision-language models with instruction tuning[J]. arXiv preprint arXiv:2305.06500, 2023.



LLaVa



VideochatGPT



Model	MSVD-QA		MSRVTT-QA		TGIF-QA		Activity Net-QA	
	Accuracy	Score	Accuracy	Score	Accuracy	Score	Accuracy	Score
FrozenBiLM	32.2	–	16.8	–	41.0	–	24.7	–
Video Chat	56.3	2.8	45.0	2.5	34.4	2.3	26.5	2.2
Video-ChatGPT	64.9	3.3	49.3	2.8	51.4	3.0	35.2	2.7

Table 2: **Zeroshot question-answering** comparison of Video-ChatGPT with other video generative models. Video-ChatGPT performs competitively across all datasets.

Evaluation:

1. Zero-Shot Video QA(OE)
2. Scored by GPT-4, including aspects like information accuracy and event time sequence.

Model:

1. The model is Llava-based, fine-tuned with a CLIP visual encoder and Vicuna language decoder, both with frozen weights.
2. CLIP features from each video frame undergo temporal and spatial pooling, then are combined to create spatiotemporal features.
3. A trainable linear layer is added after these visual spatiotemporal features.
4. Training utilizes VideoQA. LLM is the decision model.

Dataset:

1. Frame-level captions are created with BLIP, detailed subtitles by GRiT, and key frames are tagged by Tag2Text.
2. Any mismatched captions from BLIP-2 or GRiT, compared to Tag2Text tags, are eliminated.
3. GPT assembles captions into a seamless video caption.

More Test about V-L Model

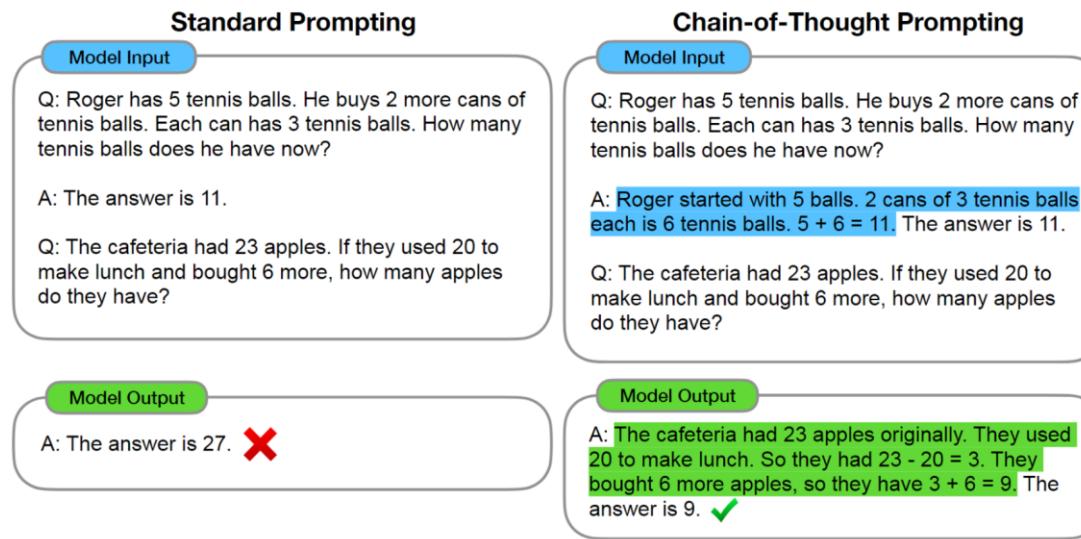
Reasoning Ability? (CoT)

Visual Encoder?

Language Model?

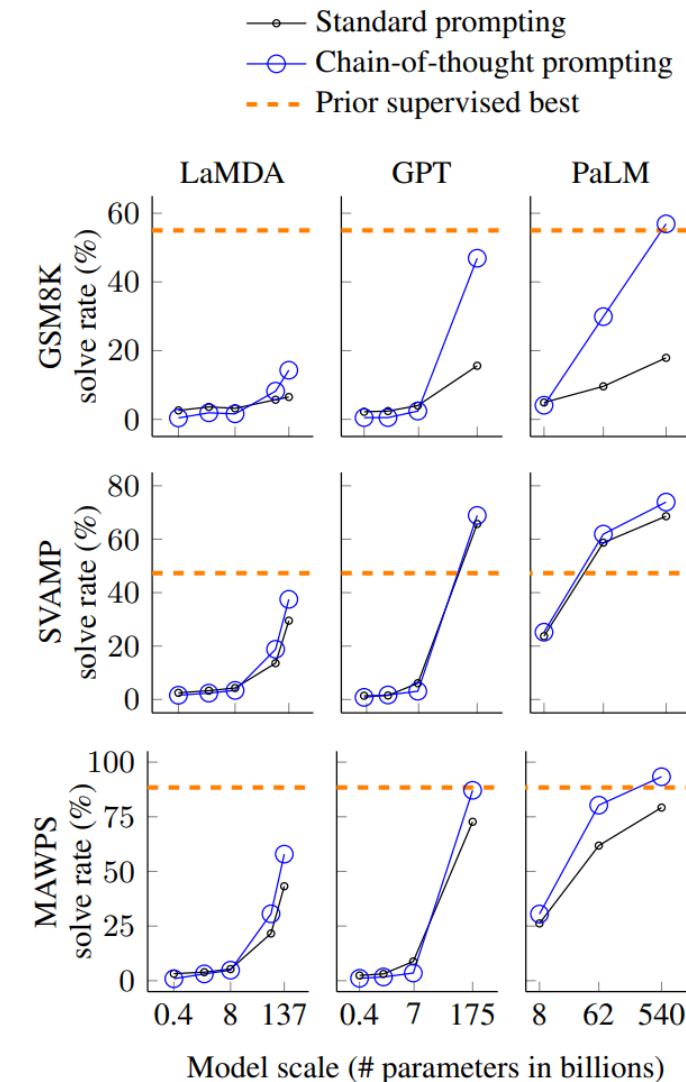
Why are LLMs Used as Decision Models in Recent Research?

[Back](#)



The strongest results of chain-of-thought prompting are summarized in Figure 4, with all experimental outputs for each model collection, model size, and benchmark shown in Table 2 in the Appendix. There are three key takeaways. First, Figure 4 shows that chain-of-thought prompting is an emergent ability of model scale (Wei et al., 2022b). That is, chain-of-thought prompting does not positively impact performance for small models, and only yields performance gains when used with models of $\sim 100B$ parameters. We qualitatively found that models of smaller scale produced fluent but illogical chains of thought, leading to lower performance than standard prompting.

1. Thought-chain prompting does not positively impact the performance of smaller models. These smaller models produce fluent but illogical thought-chains, leading to performance below that of standard prompts.
2. Thought-chain prompting brings greater performance improvement for more complex problems, while the performance enhancement for simpler problems is either negative or very small.



LLaVa Experiments

Question: why does the man playing the white guitar go near the man playing the keytar?

Let's think

step by step, According to the given question and common sense, please choose the most likely intention of the protagonist in the question from the following options.

Options:

- (0). want to talk to him
- (1). make music
- (2). to energise him
- (3). want to hit him
- (4). to annoy him



The most likely intention of the man playing the white guitar approaching the man playing the keytar is to **make music together**. Both musicians are holding their instruments and are on stage, which suggests that they are part of a performance or a jam session. As

why does the man playing the white guitar go near the man playing the keytar ?

LLaVa Experiments



The image depicts a lively scene of a group of people playing musical instruments in a living room. There are at least four people in the room, with two of them playing guitars, one playing a keyboard, and another possibly playing a different instrument. The group appears to be having fun and enjoying their time together as they perform.

The living room is furnished with a couch, a chair, and a potted plant. The couch is situated in the background, while the chair is placed closer to the center of the room. The potted plant is located near the left side of the room, adding a touch of greenery to the space.

why does the man playing the white guitar go near the man playing the keytar?

While I cannot determine the exact reason for the man playing the white guitar going near the man playing the keytar, it is likely that they are collaborating on a musical piece or engaging in a friendly interaction during their performance. Musicians often move around on stage or in a performance space to interact with their fellow musicians, adjust their instruments, or simply engage in the energy of the performance. In this case, the man playing the white guitar may be moving near the man playing the keytar to harmonize with him, exchange a musical idea, or simply to create a more dynamic and engaging performance for the audience.

choose from: want to talk to him, make music, to energise him, want to hit him, to annoy him

Based on the image and the context provided, the most likely reason for the man playing the white guitar going near the man playing the keytar is to make music together and energize each other during their performance. It is evident that they are enjoying their time and collaborating on a musical piece, which is a

LLaVa Experiments



Question: why does the boy cross his legs on the bed?

Let's think step by step, please choose the most likely answer from the following options according to the given question and common sense. options:

- (0). resting
- (1). put keyboard on his lap
- (2). wants to get on the bed
- (3). point to the video recorder
- (4). follow the instructions



The most likely reason for the boy crossing his legs while sitting on the bed is that he is trying to [find a comfortable and relaxed position to play the keyboard.](#) The image shows him sitting on the bed with a keyboard in his lap, which suggests that he is playing or



why does the boy cross his legs on the bed?

LLaVa Experiments

[Back](#)

Question: why does the boy cross his legs on the bed?



The boy is likely crossing his legs on the bed as a comfortable and relaxed position to play his keyboard. Sitting in this position can provide stability and support for his hands while playing the instrument. Additionally, it allows him to maintain a sense of comfort and ease, which can enhance his focus and enjoyment while playing the keyboard.

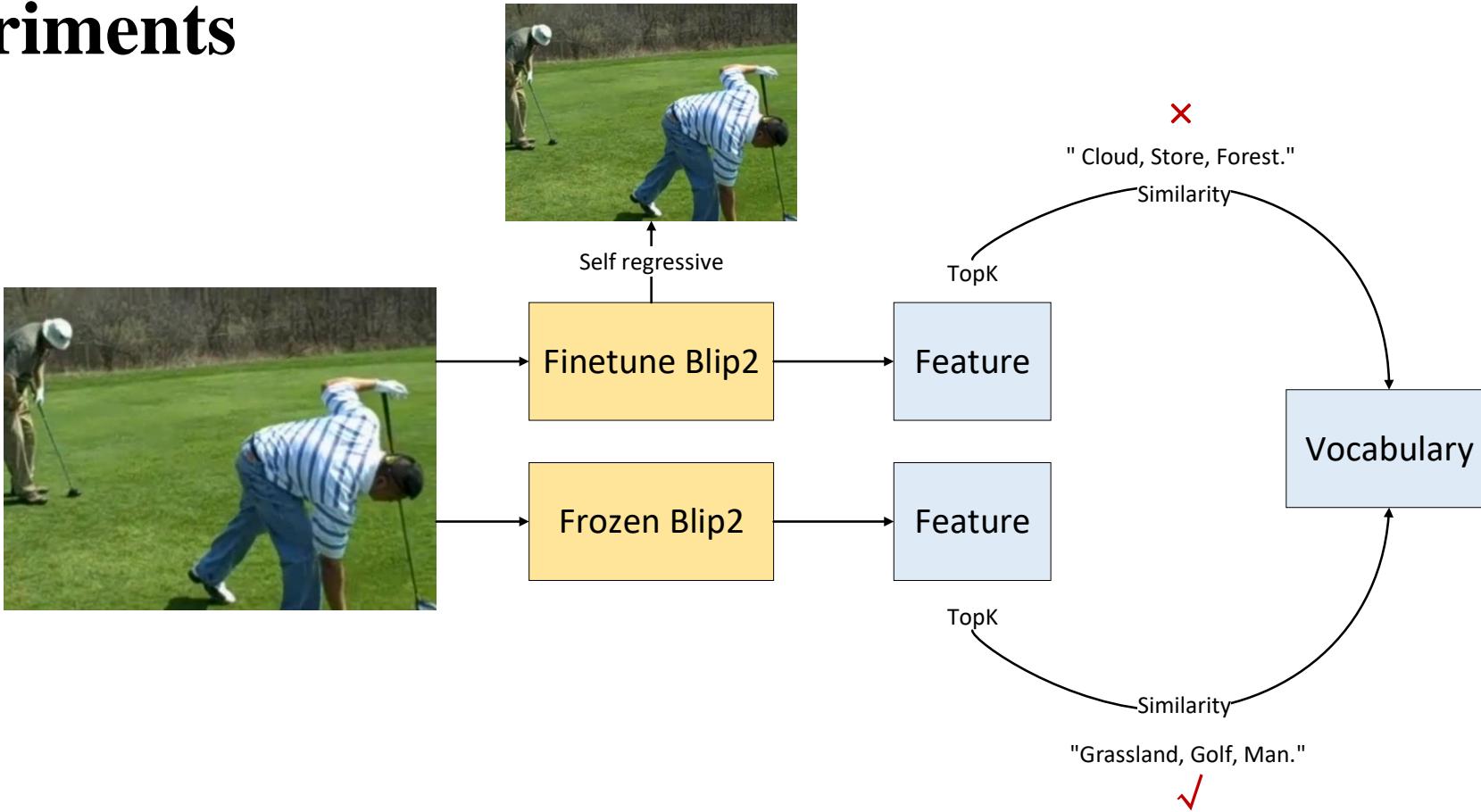
please choose the most likely answer from the following options according to the given question and common sense. options:

- (0). resting
- (1). put keyboard on his lap
- (2). wants to get on the bed
- (3). point to the video recorder
- (4). follow the instructions

The most likely answer is (1). The boy is putting the keyboard on his lap while sitting on the bed. This position allows him to play the keyboard comfortably and maintain a relaxed posture.

1. The capability of Cot is crucial. In the examples I tried, as long as I asked in multiple steps, Llava could answer correctly where GPT failed.
2. Though it's an image model and can correctly answer when provided with key frames, its video handling ability remains unproven.
3. Delaying questioning and option provision allows for more comprehensive image understanding, preserving valuable information that might be lost due to initial queries or options, given the seeming unreliability of the vision encoder's attention ability.

Blip2 Experiments



- 1.Using a trained Blip2 encoder for image feature extraction and similarity analysis with the vocabulary can highlight correspondences with the visual target, but with significant information loss.
- 2.Fine-tuning Blip2 on an image reconstruction dataset and training the vision encoder autoregressively may result in poor correspondence between image features and the vocabulary, but preserves ample information for image reconstruction.

[Back](#)

Don't These V-L Models Know that Stepping on a Scale is a Punchline?

[Back](#)

LLAVA:

The person on the left may be seriously measuring their weight while the person (Obama) behind him is laughing at the reading, causing the person on the left to believe they weigh more than they actually do.

Yes, that is a possible interpretation of the scene. The person on the left is seriously measuring their weight while Obama, in a playful manner, adds to the scale's reading, making it appear heavier than it actually is. This creates a humorous situation where the person on the left believes they weigh more than they actually do.



GPT4:

ZH 想象一幅图像。图像的左边一个人站着秤，并且在笑。图像的右侧还有一面镜子，反射出另一个人的笑容，以及图像中的两个人为什么笑。

这幅图像的幽默之处在于它表达了一种对比：左侧的人严肃地测量体重，而身后的人却恶作剧地加重了秤的读数。这种对比戏弄使得站在身后的人和站在右边的人感到困惑，从而产生一种滑稽的情况和轻微的尴尬的幽默感。

需要注意的是，这种理解依赖于观看图像的人对于社会行为规范和人类交际的基本理解。例如，我们知道人们通常在测量体重时希望获得准确的读数，我们也知道人们经常会为了娱乐而进行轻微的恶作剧。此外，人们也普遍能理解轻松和愉快的情绪会使人发笑。所以，尽管图像本身并没有文字说明，我们还是可以从中理解出这些信息。

GPT4, With Noisy Caption:

ZH 想象一幅图像。图像的左边是一个镜子，镜子里反射着右边的人。图像中间一个笑着的人向前伸出脚踩着秤，图像左侧一个人站在秤上，图像右侧一个人也在笑着，三个人的身体都朝向白色，天花板上的灯是亮着的。请从中找出造成这张图片的原因，以及图像中的两个发笑的人发笑的原因？

背景的理解，因为不同的人可能会从不同的角度找到幽默，从而判断出可能的幽默元素。

发笑，因为右侧的人的镜像被反射到镜子中，这可能会使得场景看起来比实际更为复杂。

元素，因为在秤上，可能正在查看他的体重，这在某些情况下是不寻常或出人意料的。另外，他还在笑，这可能意味着他对自己的体重情况很滑稽。

另一个幽默元素，因为这可能让观察者感到困惑，不清楚他为什么会在秤上，这可能会引发笑声。

取决于他们在场景中的角色和他们的理解。一个可能的解释是，左侧的人可能在秤上，而右侧的人可能在秤下，这很滑稽，因为它包含了一些不寻常或者出人意料的元素，比如秤可能也在互相开玩笑，或者在享受这个有趣的场景。

，实际的幽默元素和笑点可能会根据具体的视觉呈现和个人的幽默感觉而有所不同。

最有可能的笑点可能在于踩秤的人。他笑着踩着秤，这可能是因为他发现自己的体重比预期的轻，或者他在正常认为是私人的场合（测量体重）中找到了幽默感。另外两个人也可能因此而笑，尤其是如果他们知道踩秤的人平时很在意体重，现在却公开测量并对结果感到满意或惊讶。然而，这都是猜测，真正的笑点可能会根据具体的视觉呈现和个人的幽默感觉而有所不同。

Our Thinking and Our Idea

Our Thinking

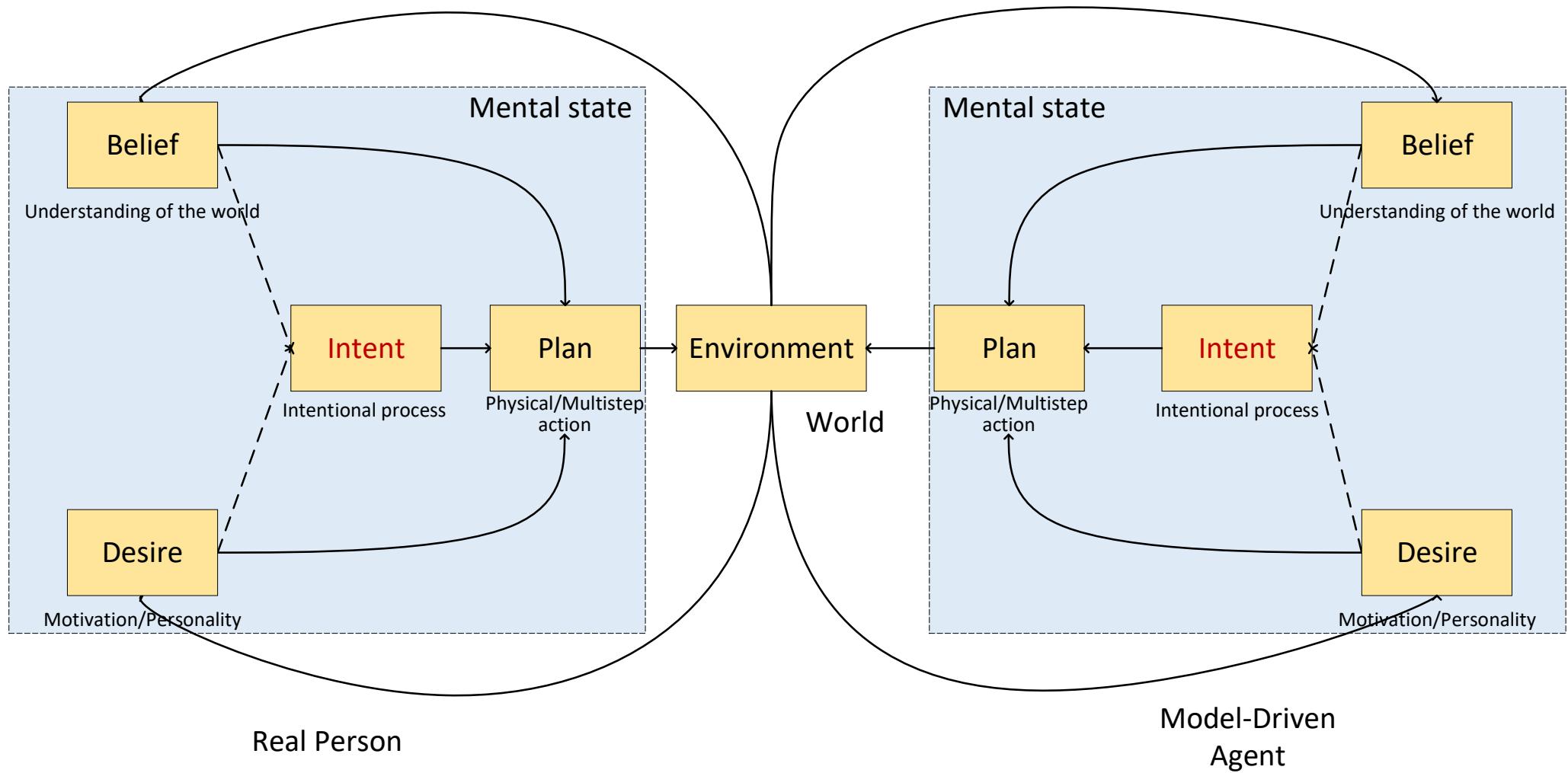
- I. The cognitive capacity of current V-L (Vision-Language) models actually lies in the **bottleneck of the visual part**. The visual part itself can't successfully use Theory of Mind (ToM) to pay attention to key visual elements, it may lose a lot of information during the encoding process.
- II. The **alignment ability** of current V-L models is poor. In the example of Obama, even if the language part knows the punchline about stepping on the scale, it can't guide the visual part of the model to discover this point in the image, thus obtaining a description of "stepping on the scale".
- III. Perhaps we could fine-tune it on a Specialized dataset for predicting humor/intention, but we might end up with a zombie that sees humor/intention in everything. **Judging when to use ToM**, and when to just judge basic visual elements, is also an important ToM ability.

Our Goals and Guiding Points

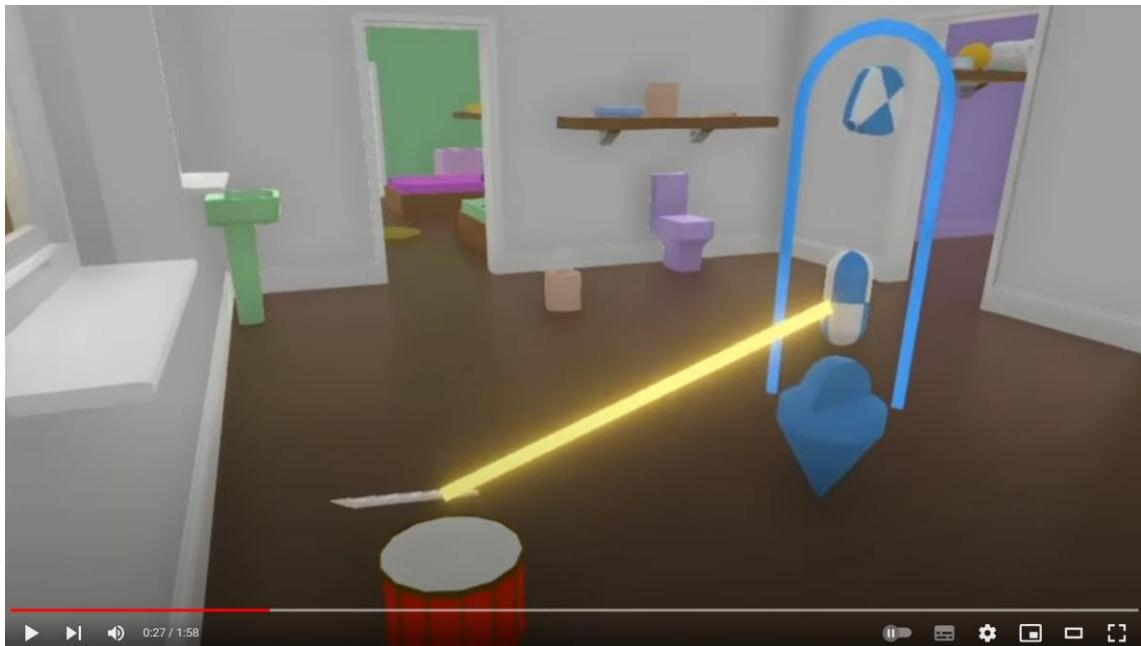
- I. We aim for a V-L model capable of determining when to use ToM skills, excelling at both videoQA and zero-shot IntentQA tasks. ([ToM test](#))
- II. Our model will utilize existing LLMs as decision models, harnessing in-context learning and Cot capabilities to improve decision-making and interpretability. ([code](#), [Cot](#))
- III. We seek to enhance vision encoder fine-tuning for efficient visual information retention and better language alignment. ([Blip2 Experiment](#))
- IV. Multiple dialogue cycles are necessary for repetitive visual information extraction, with varying queries providing diverse attention, thereby delivering comprehensive visual information to the LLM. ([LLaVa Experiment](#))

Future Work: Building a Complete ToM Process in VL Models

ToM (Theory of Mind)

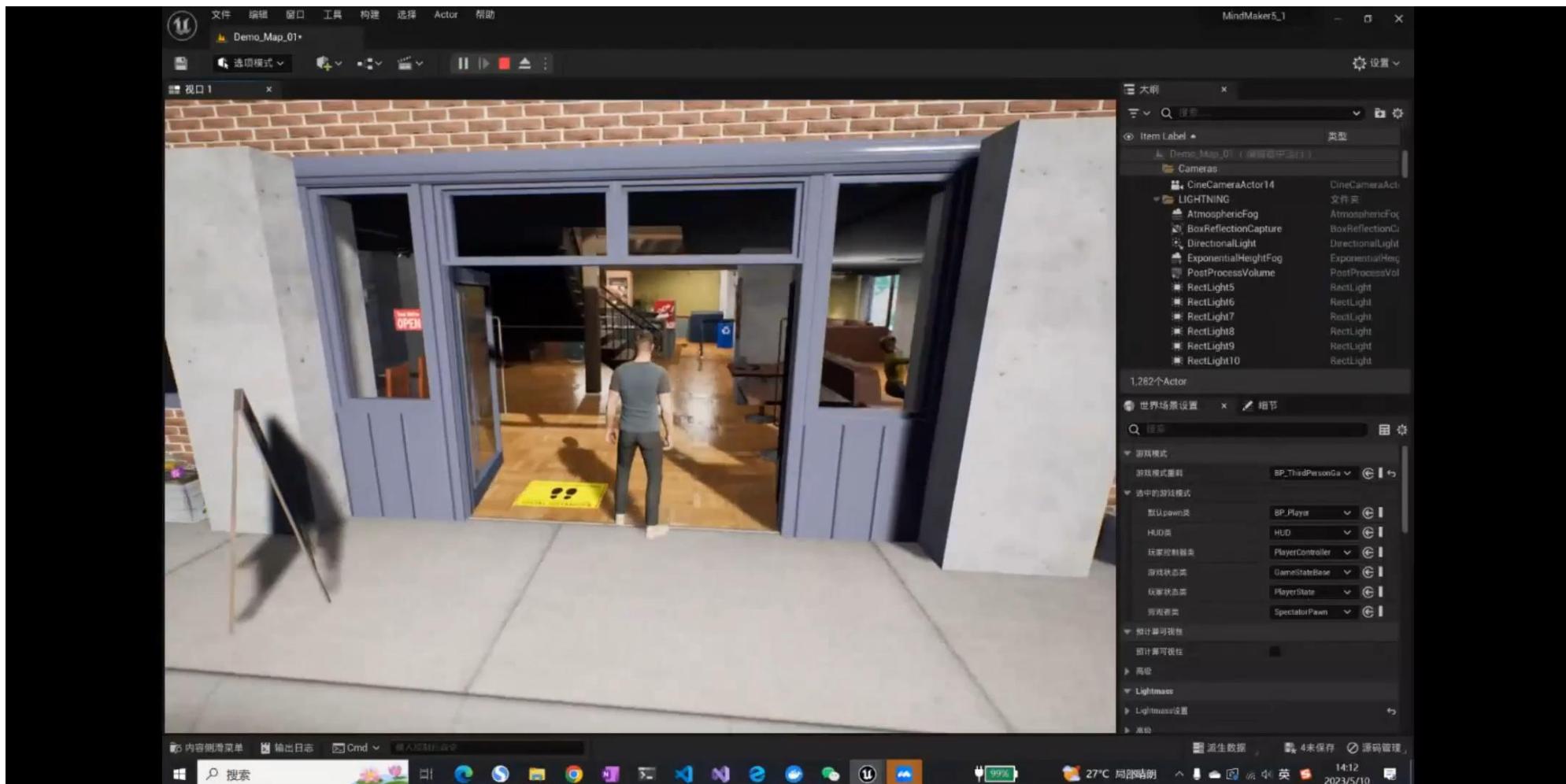


Future Work



<https://www.youtube.com/watch?v=ZFgRhviF7mY>

Future Work



Reference

- Li J, Li D, Savarese S, et al. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models[J]. arXiv preprint arXiv:2301.12597, 2023.
- Liu H, Li C, Wu Q, et al. Visual instruction tuning[J]. arXiv preprint arXiv:2304.08485, 2023.
- Liangke Gui, Borui Wang, Qiuyuan Huang, Alex Hauptmann, Yonatan Bisk, and Jianfeng Gao. Kat: A knowledge augmented transformer for vision-and-language. NAACL, 2021.
- hengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. An empirical study of gpt-3 for few-shot knowledge-based vqa. In AAAI, pages 3081–3089, 2022
- uanze Lin, Yujia Xie, Dongdong Chen, Yichong Xu, Chenguang Ma, Maaz M, Rasheed H, Khan S, et al. Video-ChatGPT: Towards Zhu, and Lu Yuan. REVIVE: Regional visual representation Detailed Video Understanding via Large Vision and Language matters in knowledge-based visual question answering. In Models[J]. arXiv preprint arXiv:2306.05424, 2023.
- NeurIPS, 2022
- Zhu D, Chen J, Shen X, et al. Minigpt-4: Enhancing vision-language understanding with advanced large language models[J]. arXiv preprint arXiv:2304.10592, 2023.
- Team D M I A, Abramson J, Ahuja A, et al. Creating multimodal interactive agents with imitation and self-supervised learning[J]. arXiv preprint arXiv:2112.03763, 2021.
- Yang Z, Li L, Wang J, et al. Mm-react: Prompting chatgpt for multimodal reasoning and action[J]. arXiv preprint arXiv:2303.11381, 2023.
- Shen Y, Song K, Tan X, et al. Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface[J]. arXiv preprint arXiv:2303.17580, 2023.
- Surís D, Menon S, Vondrick C. Vipergpt: Visual inference via python execution for reasoning[J]. arXiv preprint arXiv:2303.08128, 2023.
- Wu C, Yin S, Qi W, et al. Visual chatgpt: Talking, drawing and editing with visual foundation models[J]. arXiv preprint arXiv:2303.04671, 2023.
- Dai W, Li J, Li D, et al. Instructblip: Towards general-purpose vision-language models with instruction tuning[J]. arXiv preprint arXiv:2305.06500, 2023.
- Wei J, Wang X, Schuurmans D, et al. Chain of thought prompting elicits reasoning in large language models[J]. arXiv preprint arXiv:2201.11903, 2022.