Visual Chain-of-Thought Prompting for Knowledge-Based Visual Reasoning

Zhenfang Chen[1*], Qinhong Zhou[2*], Yikang Shen[1], Yining Hong[3], Zhiqing Sun[4], Dan Gutfreund[1], Chuang Gan[1,2]

[1]MIT-IBM Watson AI Lab
[2]UMass Amherst
[3]University of California, Los Angeles
[4]Carnegie Mellon University

Abstract

Knowledge-based visual reasoning remains a daunting task since it not only requires machines to interpret the concepts and relationships from visual scenes but also associate them with external world knowledge to conduct a chain of reasoning on open-world questions. Previous works, however, treat visual perception and language-based reasoning as two independent modules, failing to attend to both modules throughout all stages of reasoning. To this end, we propose Visual Chain-of-thought Prompting (VCTP) for knowledge-based reasoning, which involves the interaction between visual content and natural language in an iterative step-by-step reasoning manner. VCTP contains three stages, **see**, **think** and **confirm**. The **see** stage scans the image and grounds the visual concept candidates with a visual perception model. The **think** stage adopts a pre-trained large language model (LLM) to attend to key visual concepts from natural language questions adaptively. It then transforms key visual context into text context for prompting with a visual captioning model, and adopts the LLM to generate the answer. The **confirm** stage further uses the LLM to generate the supporting rationale for the answer, which is then passed through a cross-modality classifier to verify that it's consistent with the visual context. We iterate through the think-confirm stages to ensure the verified rationale is consistent with the answer. We conduct experiments on a range of knowledge-based visual reasoning datasets. We found our VCTP enjoys several benefits, 1). it achieves better performance than the previous few-shot learning baselines; 2). it enjoys the total transparency and trustworthiness of the whole reasoning process by providing rationales for each reasoning step; 3). it is computation-efficient compared with other fine-tuning baselines. Our code is available at https://github.com/UMass-Foundation-Model/VisualCoT.git

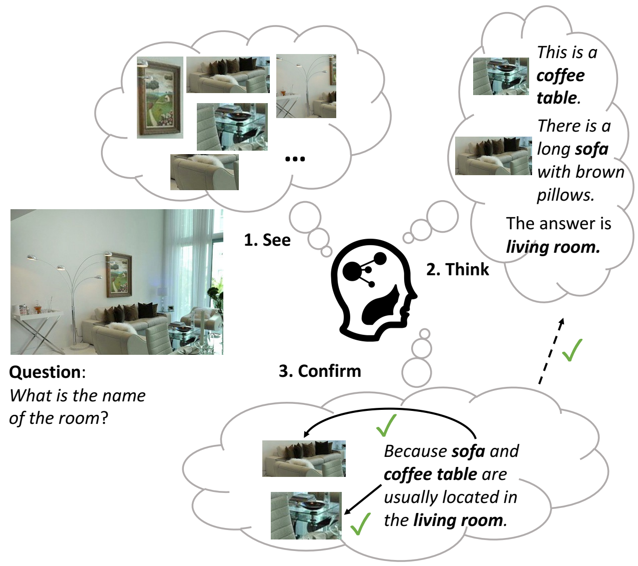

Figure 1: The human process to handle knowledge-based visual reasoning. Given an image-question pair, a human is able to *see* all the objects in the image, *think* of the related visual concepts to get the answer, and finally *confirm* the answer is correct based on visual observation and knowledge.

Introduction

Machine visual reasoning has taken a huge step forward since neuro-symbolic mechanisms (Yi et al. 2018) were introduced, which enable machines to develop a reasoning chain with multiple steps. However, as foreshadowed by cognitive scientists in early works, such systems of logic and symbols are fundamentally ill-suited to representation and reasoning with real-world, common-sense knowledge (Oaksford and Chater 2007) since they rely solely on closed-world logic rules and hard constraints(Stenning and Van Lambalgen 2012).

We therefore study the problem of knowledge-based visual reasoning (Marino et al. 2019; Schwenk et al. 2022), which requires models to interpret the image content, recall the relevant parts of open-world knowledge, and perform step-by-step logical reasoning to arrive at an answer. Knowledge-based visual reasoning is more challenging than traditional visual question answering and reasoning (Antol et al. 2015; Goyal et al. 2017) in that the visual context, external knowledge, and natural language questions need to be interactively integrated throughout the reasoning chain. As depicted in Fig 1, to answer the question *"What is the name of the room?"*, humans need first to ***see*** the room and extract visual concepts such as *"frame"*, *"sofa"*, and *"lamp"*. We then attend to key visual concepts that are semantically related to the question and ***think*** that *"This is a coffee table"* and

*indicates equal contributions

"There is a long sofa with brown pillows" to get the answer *"living room"*. Last but not least, humans would finally ***confirm*** the answer is correct by calling back to the visual context and drawing the conclusion *"Sofa and coffee table are usually located in the living room"*. The vision-to-language and language-to-vision interactions can be iteratively performed until we arrive at a satisfying answer. If machines are also endowed with such capabilities, they can be utilized for numerous real-world applications such as assistive robots (Brohan et al. 2022), and embodied chatbots (Konstantopoulos 2010).

Consistent with cognitive scientists' notion that knowledge with soft constraints is eminently compatible with large connectionist models (Oaksford and Chater 2007), large language models (LLMs) indeed have made tremendous progress in few-shot reasoning (Brown et al. 2020; Zhang et al. 2022). In particular, chain-of-thought prompting (Wei et al. 2022) transcends vanilla LLMs by providing a chain of thought (*i.e.*, a series of intermediate reasoning steps) to perform complex language reasoning. However, it remains a challenge how to leverage these LLMs for reasoning in vision and language tasks. Existing works either finetune the LLMs with massive vision-language data (Tsimpoukelli et al. 2021; Alayrac et al. 2022; Jin et al. 2022), which is computational-intensive; or adopt prompt-based methods such as PICa (Yang et al. 2022) to translate images into captions for textual prompting. Both lines of work are incapable of generating step-by-step reasoning chains like the reasoning process of human beings, leaving the models completely black boxes. Furthermore, they treat visual perception and language-based reasoning as independent modules (*e.g.*, the visual modules are used for generating features or captions for LLMs), neglecting the massive *interaction* and *communication* across modalities. This deviates from the way human beings perform visual reasoning as in Figure 1. Inspired by this, the key question we would like to investigate is how we can large language models to interact and communicate with visual information to construct step-by-step reasoning chains for open-world knowledge-based visual reasoning. To this end, we propose a novel framework named Visual Chain-of-Thought Prompting (VCTP) to mimic the human reasoning process in Fig. 1. Our model is an iterative and interactive framework with three key modules, a ***see*** module, a ***think*** module, and a ***confirm*** module. Given an image-question pair, the ***see*** module first uses a scene parser to extract all the candidate visual concepts in the image. The ***think*** module adopts an LLM to select relevant visual concepts (*e.g. "Sofa"* in Fig. 1) extracted by the ***see*** module corresponding to the given natural language question and uses a captioning model to transform visual information into textual descriptions(*e.g. "There is a long sofa with brown pillows"*). The LLM predicts the answer to the question (*"living room"*) based on the attended visual context. Moreover, we introduce a ***confirm*** module for rationale verification to provide more transparent and trustworthy reasoning. Specifically, we require the LLM to generate rationales (*e.g. "Sofa and coffee table are usually located in the living room"* for the predicted answer in Fig. 1). We then estimate the matching similarity between these rationales and the given visual input with a neural cross-modality classifier. Finally, the selected rationale is fed to the LLM's prompt to ensure that the rationale can infer the same output consistently. We repeat the process of ***think*** and ***confirm*** iteratively for sufficient cross-modality interactions until the answers from two consequent iterations are the same.

To summarize, we introduce VCTP, a novel modularized, interactive, and iterative framework for knowledge-based visual reasoning, which is able to iteratively attend to the related visual concepts in the image and provide consistent supporting rationales for the answer prediction. VCTP enjoys several advantages. First, it is effective. Extensive experiments on knowledge-based benchmarks found that VCTP achieves better performance than previous few-shot baselines. Moreover, VCTP is more transparent and interpretable since it maintains the whole step-by-step reasoning trace that leads to the prediction. VCTP is also more computationally efficient compared with finetuning methods. Our code would be available on acceptance.

Related Work

Visual Prompting. Large language models like GPT-3 (Brown et al. 2020) have popularized few-shot prompting in natural language processing (NLP), where several input-output pairs are used as context for the language model to understand the task and generate predictions for a new example. Chain-of-thought prompting (Wei et al. 2022) and its variants (Creswell, Shanahan, and Higgins 2022; Zhou et al. 2022a; Marasovic et al. 2022) have been developed for more effective or transparent reasoning in NLP. Later, prompting was brought to the vision community (Zhou et al. 2022c; Jia et al. 2022; Ju et al. 2021; Ge et al. 2022; Zhou et al. 2022b; Wang et al. 2022). CLIP (Radford et al. 2021) and Region-CLIP (Zhong et al. 2022) enable zero-shot classification and detection by replacing the class labels with natural language supervision during training. UnitedIO (Lu et al. 2022) specifies each vision task with a language prompt to perform multi-task learning. Concurrent work (Zhang et al. 2023) trains a transformer to predict rationales first and then infers the answer based on multimodal feature vectors. Differently, we aim to use interactive prompting for knowledge-based visual reasoning. It requires frequent interaction between language models and vision models, such as extracting related visual concepts, recalling external knowledge, and verifying the text prediction consistent with the visual context, which has not been well studied before.

Large Pre-trained Models for Visual Reasoning. Large pre-trained models have also been used in reasoning over vision and language (Dou et al. 2022; Zeng et al. 2022; Gao et al. 2022; Zhao et al. 2023). Most works (Tsimpoukelli et al. 2021; Jin et al. 2022; Zellers et al. 2022; Chen et al. 2022a) learn large pre-trained models from massive vision-language data and finetune them for downstream tasks, which are usually extremely computationally expensive and time-consuming. For example, Flamingo (Alayrac et al. 2022) needs to be finetuned on 1536 TPUv4 for 15 days with 185 million images and 182 GB of text. It has also been observed that many of these pre-trained models (*e.g.*, (Kamath et al. 2022; Tan and Bansal 2019)) contain limited open-world knowledge. They achieve inferior performance on KB-VR datasets, compared with models with LLMs for

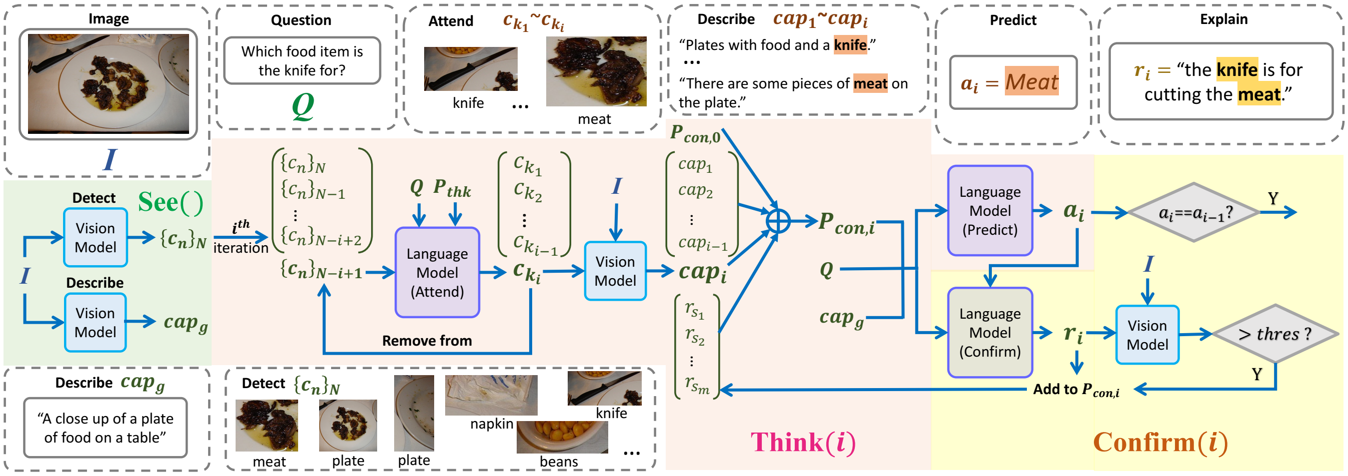

Figure 2: The framework of our VCTP. Given an image-question pair, we first use the *see* module to detect all object candidates in the image and translate the whole image into a global description. Then, the *think* module adopts an LLM to attend to the key visual concepts, transforms the selected concept into a language description with a captioner, and leverages the LLM to predict an answer. The *confirm* module requires the LLM to continue the generate the supporting rationale, verify whether the rationale is consistent with the image content, and ensure that the same answer can be produced when the rationale is added to the prompt in the next iteration. Algorithm 1 provides a detailed description of the VCTP framework.

external knowledge (Gui et al. 2022). Yang *et al*. converted images into textual descriptions and treated them as prompts for LLMs, which achieves high performance on knowledge-based visual question answering. However, its text-based visual context is independent of the query question and leaves the question-answering process a black box.

Knowledge-based Visual Reasoning. Our work is also related to knowledge-based visual reasoning (KB-VR) (Marino et al. 2019; Schwenk et al. 2022; Wang et al. 2017a), which requires both understanding the image content and retrieving external knowledge to answer the questions. Most early methods (Zhu et al. 2021; Ding et al. 2022b; Gardères et al. 2020; Lin and Byrne 2022) use deep neural networks to understand images and retrieve relevant knowledge from explicit knowledge bases. Recent methods (Gui et al. 2022; Yang et al. 2022) found that LLMs like GPT-3 could serve as a knowledge base and use the LLM to answer the question directly. While our model also retrieves relevant knowledge from LLMs, it provides the step-by-step reasoning process besides the final answer prediction.

Neural-Symbolic Visual Reasoning. Our work can be regarded as neural module networks (Yi et al. 2018; Andreas et al. 2016; Chen et al. 2022c; Ding et al. 2021, 2022a; Chen et al. 2021b), which provides transparent step-by-step reasoning. They typically decompose the query question into a set of operations, model each operation with a network module, and iteratively execute these operations to get results. While these models are more interpretable, they either focus on simple physical scenes (Johnson et al. 2017; Yi et al. 2020) or only achieve inferior performance on real-world datasets (Hudson and Manning 2019), compared with end-to-end neural network methods (Chen et al. 2022b). Different from them, we would like to build a system that can achieve reasonable performance on open-world knowledge-based visual reasoning while maintaining the system's interoperability.

Method

Overall

In this section, we introduce a novel framework called Visual Chain-of-Thought Prompting (VCTP) for knowledge-based visual reasoning, which can understand the query question, attend to key visual concepts in the image, retrieve supporting evidence, and finally get the answer in a step-by-step manner. VCTP consists of three modules, ***see***, ***think*** and ***confirm***, and run these modules in an iterative manner. As illustrated in Fig. 2, given an image and a question about its content, the ***see*** module uses a scene parser (Han et al. 2021) to detect all the candidate objects (concepts) in the image and represents them with their predicted class names. It also generates a global description for the whole image. Then, ***think*** module attends to the key concepts that are semantically related to the query question with an LLM and describes them in the form of natural language with an image captioner. Based on the attended visual context, the LLM predicts the answer to the question. The ***confirm*** module requires the LLM to continue to generate the answer's supporting rationale and verify them with a cross-modality classifier (Radford et al. 2021). To ensure the rationale is consistent with the answer, we add the verified supporting rationale back into the prompting context and begin a new ***think-confirm*** iteration. We iteratively generate the answer and the rationale until the answer predictions in two consequent iterations are consistent. We summarize the whole algorithm flow in Algorithm 1.

Compared with existing learning-in-context methods (Yang et al. 2022; Wei et al. 2022), our framework has two advantages, effectiveness and interpretability. It is effective since it can adaptively attend to the related visual regions and output consistent answer-rationale predictions. It is interpretable as it is able to perform a step-by-step investigation of the whole reasoning process. Visualized examples of such a reasoning process can be found in Fig. 4.

Model Details

See Module. Given a query image, we use a Faster-RCNN (Ren et al. 2015) to detect all the object candidates in the image. Specifically, we use the detection model released by Yang *et al*. (Han et al. 2021) to predict object locations and their categorical labels such as *"knife"*, *"plate"* and *"napkin"*. It also provides a global caption for the whole image with an image captioner (Li et al. 2022). These visual concepts will be selected and further described in the ***think*** module to provide valuable visual context to get the answer.

Think module. The second module of our framework is the ***think*** module, which attends to the corresponding regions in the image and transforms them into the textual description for the LLM to predict the answer. We use an *attend-describe-predict* approach in the ***think*** module, as shown in Fig. 2.

The first step is *attend*, where we use prompting methods (Brown et al. 2020; Chowdhery et al. 2022) to help the LLM to attend to the key concepts in the image that is semantically related to the query question. We show the prompting template for the LLM to attend to key visual concepts in Fig. 3 (A). We feed some input-output pairs from the training set into the LLM's prompt and ask the LLM to select based on the given context. The question is shown on the top of the template. Objects detected in the ***see*** module are represented by their category names, such as *"knife"* and *"beans"*, and the vocabulary of LLM output words is constrained to these category names. The LLM selects the most related object to provide further visual context to handle the query question. As shown in Fig. 2, such attended concepts (*e.g. "meat"* and *"knife"*) are important to get the answer *"meat"* to the question *"Which food item is the knife for?"*.

The next step of the ***think*** module is *describe*. The region in the image corresponding to the selected concept is cropped and fed to an image captioner (Li et al. 2022) to generate a regional description for the new concept. The generated regional descriptions will be added to the LLM's prompt to provide fine-grained visual context to predict an answer. Note that a regional description for an object is usually more informative than the object class. For example, the caption of *"some pieces of meat on the plate"* in Fig. 2 additionally describes the relationship between *"meat"* and *"plate"*.

The last step of the ***think*** module is *predict*. We add multiple question-answering examples from the training set to the LLM's prompt and ask it to predict an answer. The answer prediction is based on the attended visual context and the rationale predicted by the ***confirm*** module in the previous iterations, which we will discuss in the ***confirm*** module.

Confirm Module. The last module of our framework is the ***confirm*** module, as shown in Fig. 2, which aims to generate a consistent supporting rationale for the answer prediction and verify the prediction's correctness. Given the few-shot example context and the interactive prompt generated by the ***think*** module, we require the LLM to continue to predict the supporting rationale after the answer prediction. A prompt

Algorithm 1: Pipeline of the proposed VCTP

Input: Input Image and Question $\{\mathbf{I}, \mathbf{Q}\}$.
Output: Answer and the reasoning process $\{\mathbf{a}^*, \mathbf{R}\}$
Require: c_n is the n-th concept detected in the image; c_{k_i} and cap_i are the concept and the regional caption at the i-th iteration; cap_g denotes a caption for the global image. $mIter$: the maximal iteration; P_{thk} and P_{con} are the prompt for concept selection and question answering.
1: # the ***see*** module
2: $\{c_n\}_N \leftarrow$ `ImageParser` $(\mathbf{I})$
3: $cap_g \leftarrow$ `GlobalCaptioner` $(\mathbf{I})$
4: i starts from 0. a_0 is an empty string; $P_{con,0}$ is the in-context examples, answer, and rationales; P_{thk} is the in-context examples of the question and object selection.
5: **repeat**
6: $\quad i \leftarrow i + 1$
7: $\quad$ # the ***think*** module
8: $\quad c_{k_i} \leftarrow$ `LLMAttend` $(\{c_n\}_N \setminus \{c_{k_j}\}_{j=0}^{i-1}, \mathbf{Q}, P_{thk})$
9: $\quad cap_i \leftarrow$ `Captioner` $(c_{k_i}, \mathbf{I})$
10: $\quad P_{con,i} \leftarrow P_{con,i-1} + cap_i$
11: $\quad a_i \leftarrow$ `LLMPredict` $(P_{con,i}, \mathbf{Q})$
12: $\quad$ # the ***confirm*** module
13: $\quad r_i \leftarrow$ `LLMConfirm` $(P_{con,i}, \mathbf{Q}, a_i)$
14: $\quad$ **if** `Verify` $(r_i, \mathbf{I}) > thre$ **then**
15: $\qquad P_{con,i} \leftarrow P_{con,i} + r_i$
16: $\quad$ **end if**
17: **until** $a_i = a_{i-1}$ **or** $i = mIter$
18: $\mathbf{a}^* \leftarrow a_i$; $\quad \mathbf{R} \leftarrow (\{c_{k_j}, cap_j\}_{j=1}^{i}, r_i)$
19: **return** $\{\mathbf{a}^*, \mathbf{R}\}$

example of such a *question-answer-rationale* template can be found in Fig. 3 (B). A significant problem of the LLM's prediction is that the generation procedure is a black box, and it is difficult to verify the correctness of the predicted answer and rationale. We believe a correct rationale should have two distinct features. First, the rationale should be consistent with the answer. Given the predicted supporting rationale (*"The knife is for cutting the meat"* in Fig. 2) for the answer (*"meat"*), we should be able to predict the same answer when it is added to the context. Second, the rationale should be consistent with the visual input.

To ensure that the rationale supports the answer prediction, we feed the generated textual rationale into the LLM's prompt in the next iteration. We repeat this *answer-to-rationale* and *rationale-to-answer* procedure until the two consequent predicted answers are the same (Line 4 to Line 17 of the Algorithm 1). To ensure that the generated rationale is consistent with the given visual context, we use a large pre-trained cross-modality classifier (Radford et al. 2021) to verify whether the textual rationale matches the given images or not (Line 13 to Line 15 of the Algorithm 1). Only the rationale with the high matching similarity will be accepted and added to the prompt for the answer prediction in the next iteration.

Experiments

In this section, we demonstrate the advantages of the proposed VCTP with extensive experiments.

Implementation Details. The effectiveness of the proposed VCTP relies on the interaction of several pre-trained vision

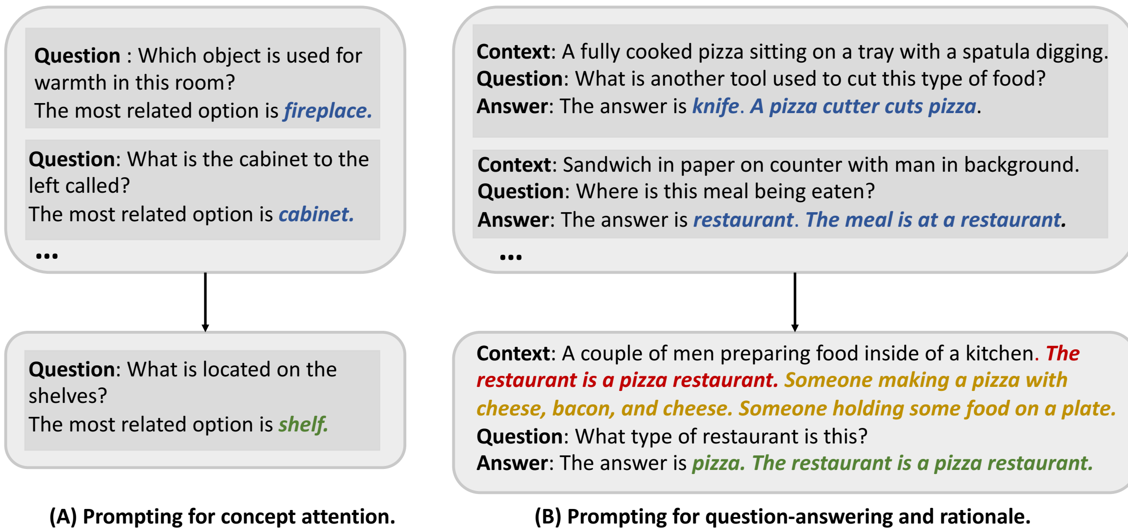

Figure 3: Prompting examples for visual concept attention (P_{thk} in Algorithm 1) and question-answer-rationale reasoning (P_{con} in Algorithm 1). Outputs in the in-context examples and test examples are marked with blue and green colors. The attentive regional captions and the rationale in the previous iterations are marked with red and bronze colors. We regard the most related option in the in-context examples as the option (concept) closest to the ground-truth answer by CLIP similarity.

models (Li et al. 2022; Radford et al. 2021) and language models (Zhang et al. 2022). We choose the faster R-CNN model (Ren et al. 2015) released by (Han et al. 2021) to detect visual concepts in images, which was trained on Visual Genome (Krishna et al. 2017). We select BLIP (Li et al. 2022) as the regional captioning model for the attended objects. We use the OPT-66B and Llama-2-70B as the pre-trained language model to prompt since they are effective and publicly available. We verify the matching similarity between the rationale and the image with the CLIP model (ViT-B/16) (Radford et al. 2021). We do not use GPT-3 (Brown et al. 2020) due to its expensive API and limited visit frequency.

We fix the number of the in-context examples in ***think*** module to 8 since it is the largest number we could efficiently run on our hardware configuration. Following (Yang et al. 2022), we prompt the LLM with in-context example selection and multi-query ensemble. For in-context examples, we select the examples most similar to the current image-question pair in training set with their clip features. For multi-query ensemble, we feed our models and the baselines 5 times and select the one with the highest log-probability as previous methods (Yang et al. 2022; Chen et al. 2021a) except the aligned models in Table 3, where we ensemble 14 times for baselines to make them have similar computation cost as ours.

Datasets and Evaluation Metric. We evaluate our models on standard KB-VR benchmarks, OK-VQA (Marino et al. 2019) and A-OKVQA (Schwenk et al. 2022). OK-VQA is the most popular knowledge-based VQA dataset with 14,055 image-question pairs. A-OKVQA is the current largest KB-VR dataset, which not only asks knowledge-related questions but also provides supporting rationales, making it a better testing bed for step-by-step reasoning. We do not conduct experiments on other KB-VR datasets like F-VQA (Wang et al. 2017a) and KB-VQA (Wang et al. 2017b), since they assume the question knowledge could be retrieved from pre-defined knowledge bases.

Baselines. We mainly compare our methods with two strong learning-in-context baselines, PICa (Yang et al. 2022) and CoT (Wei et al. 2022).

- **PICa**. PICa with GPT-3 is the current state-of-the-art few-shot model on OK-VQA, which prompts the LLM with only the image caption and object tags.
- **CoT**. CoT is a popular prompting method, performing step-by-step reasoning to solve the task rather than directly output the answer. We implement CoT by asking the LLM to generate the rationale first and then predict the answer.

We carefully implement these two methods with the OPT-66B model for a fair comparison. We also include other fully-supervised methods in the tables for performance reference.

Quantitative Results. We compare our VCTP with baselines on the validation and test sets of the A-OKVQA dataset in Table 1. Few-shot learning-in-context methods are marked with $\star$ in the table, and our method is marked with gray background for easy reference. According to the results, we have the following observations. First, we have constant gains compared with the learning-in-context baselines, PICa and CoT and even achieve better performance than the previous full-supervised pre-trained method GPV-2 on the test split of A-OKVQA. These gains show our method's effectiveness in answer predictions. We have also noticed that fully-supervised methods have a more significant performance disparity between the validation and test splits than the learning-in-context methods. For example, GPV-2 has an accuracy drop of 7.9, while our method only drops 0.4. We believe the reason is that the validation set has frequently been evaluated in fully-supervised methods and is somewhat overfitted.

We also evaluated our method in the OK-VQA in Table 1. The training set of the OK-VQA does not provide rationales for reasoning, which is needed in CoT and our method. For our method, we develop a variant that only uses examples in OK-VQA as prompting for the ***think*** module without rationale reasoning. The model stops when the predicted answers from two consecutive iterations become the same. We observe that our model performs better than the baselines, PICa and CoT, which is consistent with our finding in the A-OKVQA dataset. We can also see that (KAT-GPT-3 (Gui et al. 2022) and PICa-GPT-3) rely on much more powerful LLM, GPT-3 (175B), to achieve great performance. In contrast, our method achieves reasonable performance by integrating the available OPT-66B LLM. When we replace the OPT-66B LLM with the Llama-2-70B model (Touvron et al. 2023), or BLIP2 (Li et al. 2023) and Codex (Chen et al. 2021a) models, the performance of our method increases significantly, indicating that our method can benefit from stronger models.

Qualitative Results. The step-by-step reasoning nature of our VCTP provides better transparency and interoperability, which makes it easy to understand how our model works. We provide a qualitative comparison with baselines in Fig. 4. The $\rightarrow$ in Fig. 4 shows the flows of our method's reasoning process to get related visual concepts, regional descriptions, and the supporting rationale.

Compared with PICa, our method can adaptively attend to key visual concepts (*e.g. "ball"* and *"tennis court"* in Fig. 4) in the image that are semantically important to the question (*e.g. What is the fence meant to block?*) and describe that in the form of the natural language (*e.g. "Someone has a tennis racket and is about to hit the ball"*) to get the answer. Besides, as shown in the *explain* column of Fig. 4, our method could generate better supporting rationale (*"The wall is used for displaying art"*), which matches with the visual context in images better. In contrast, the CoT might generate a rationale inconsistent with visual context (*"The wall is used for a sofa."*), which leads to a wrong answer.

Rationale Evaluation. To further evaluate the reasoning process of our method, we compare the quality of rationales between our method and CoT in Table 2 on the validation set of A-OKVQA, where the rationales are publicly available. We use widely-used BLEU scores and CLIP sentence similarity as the metrics. We measure multi-BLEU score and averaged cosine similarity of sentence representations calculated by CLIP (ViT-B/16) (Radford et al. 2021) text encoder. Both BLEU and similarity results show that the rationales generated by our method are closer to the ground truth than the rationales generated by CoT.

Computation Analysis. To better understand the efficiency of our method, we make in-context learning baselines have similar computational costs as our model. Specifically, we increase the number of queries to ensemble k (5 in all the experiments except PICa-aligned and CoT-aligned below) for PICa and CoT, and make their overall amount of queries to LLM the same as our method. Considering the mean amount of queries is 13.62 for our method during experiments, the cost-aligned PICa and CoT have an ensemble amount $k = 14$ for each sample, denoted as PICa-aligned and CoT-aligned.

We compare their results with our method in Table 3,

Methods	A-OKVQA Val	A-OKVQA Test	OK-VQA Test
MAVEx (Wu et al. 2022)	-	-	41.37
UnifER (Guo et al. 2022)	-	-	42.13
Pythia (Yu Jiang* et al. 2018)	25.2	21.9	-
ViLBERT (Lu et al. 2019)	30.6	25.9	-
LXMERT (Tan and Bansal 2019)	30.7	25.9	-
KRISP (Marino et al. 2021)	33.7	27.1	38.4
PICa*-GPT-3	-	-	48.0
GPV-2 (Kamath et al. 2022)	**48.6**	**40.7**	-
KAT-GPT-3	-	-	**54.4**
BLIP2	38.2	37.2	45.9
CoT* (Wei et al. 2022)	41.5	43.7	38.1†
PICa* (Yang et al. 2022)	42.4	43.8	42.9
Ours*	46.4	46.0	44.6‡
Ours-Llama-2*	50.5	**54.4**	54.9
Ours-BLIP2-Codex*	**53.2**	53.8	**56.2**

Table 1: Performance comparison of our model and other baselines on A-OKVQA and OK-VQA datasets. $\star$ denotes learning-in-context methods. $\ddagger$ denotes our model's *confirm* module is removed since there are no available examples with rationales in OK-VQA. $\dagger$ denotes in-context examples with the rationales for CoT are from A-OKVQA dataset. Our model performs better than baselines.

Methods	BLEU	Sentence Similarity
CoT	14.19	80.92
Ours	**14.34**	**81.22**

Table 2: Rationale performance comparison of our model and CoT baseline on A-OKVQA validation set.

which indicates that our method still outperforms PICa and CoT significantly after considering the computational cost issue. We also notice a slight performance drop for PICa when its ensemble amount k increases. We suggest that this drop is related to the in-context sample selection method, which chooses more irrelevant samples to the current question when k increases.

Ablation Study. We conduct ablation studies on the validation set of AOK-VQA and summarize the results in Table 4. **Ours** denotes our default model with all the component integration with the $mIter$ in algorithm 1 to be 5. **Ours-1/4/6** denote that we set the maximal iteration to be 1, 4, and 6, respectively. **"w/o A/R/V"** denote the VCTP model without the *attend -describe* components, without generating the rationale and without CLIP verification module, respectively. **R-A/G/R**, respectively, attend to a random region, replace the global caption with a random one in the training set and select a random rationale from the training set; ***O-A*** is an oracle model attending to the visual concept that has the highest similarity to the answers estimated by the text encoder (Radford et al. 2021). **Recall-1** and **Recall-2** evaluate models' performance to capture key concepts, using the prompt of the last step to calculate the Recall of the ground-truth answer's most relevant visual concept and all concepts included

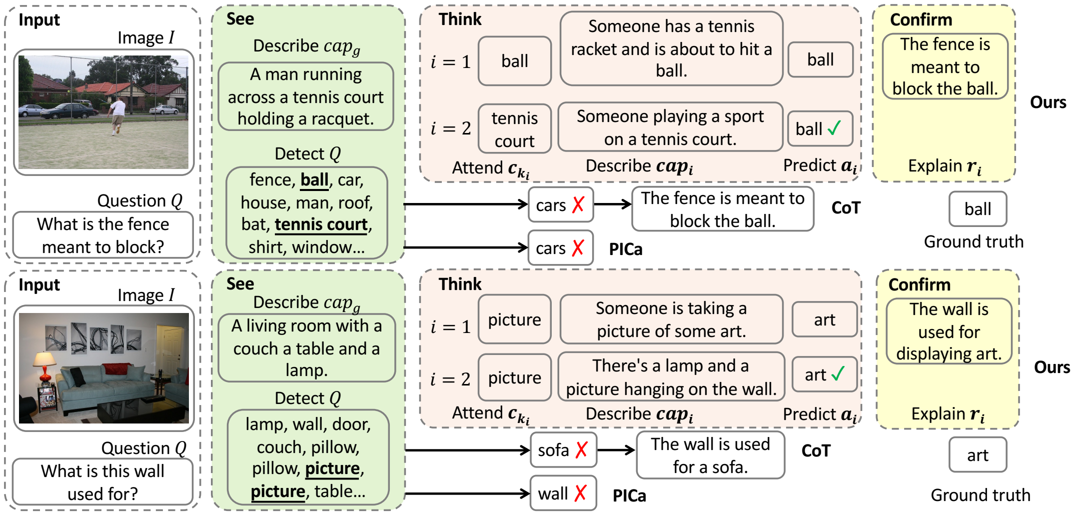

Figure 4: Qualitative results of our VCTP and baselines. Our method also enjoys better transparency by providing a step-by-step reasoning trace with related visual concepts, regional descriptions, and a supporting rationale. → denotes our reasoning flow.

Methods	A-OKVQA	OK-VQA
PICa	42.40	42.94
PICa-aligned	41.90	42.84
CoT	41.53	38.13
CoT-aligned	42.10	38.15
Ours	**46.41**	**44.62**

Table 3: Analysis of computational cost on A-OKVQA validation set and OK-VQA set. Our method still outperforms PICa and CoT after considering the computational cost.

in the ground-truth answer and rationales. We find that the performance will drop without any ablated component (**w/o A/R/V**), showing each component has its contribution to the overall performance. Compared with most ablation studies except **w/o V**, we found that our model attends to more important concepts as indicated by **Recall-1** and **Recall-2**. The model's performance could be further improved if attending to key concepts accurately (**Oracle-A**). We further find that **Ours** attends to less important concepts than **w/o V** but has higher question-answering performance, which we believe the reason is that the verification mechanism has helped rejected adding some misleading rationales back to the prompt. We observe that the *"attend"* module has the most significant effect on performance, where we think the reason is that the *"attend"* component has provided essential visual context for the LLM to infer the correct answer as shown in the examples in Figure 4. The **"w/o R"** and **"Random-R"** ablations show that adding the reasoning rationale into the model not only increases the model's transparency but also has positive effects on the answer prediction. Compared to the ablation without iteration (**Ours-1**), our model has better performance, showing the importance of iterative interactive prompting.

-	Recall-1	Recall-2	Acc.
w/o A	50.7	47.9	43.48
w/o R	64.2	52.5	45.54
w/o V	67.3	65.1	45.82
Random-A	58.7	54.5	43.21
Random-G	60.5	54.1	38.98
Random-R	65.2	53.2	44.53
Ours-1	57.9	46.8	44.68
Ours	65.9	**60.2**	46.41
Ours-4	**66.0**	59.5	**46.42**
Ours-6	65.6	60.0	46.38
Oracle-A	78.6	63.0	47.72

Table 4: Analysis on concept recall and QA accuracy.

Conclusion

In this paper, we develop a novel model named Visual Chain-of-Thought Prompting (VCTP) for knowledge-based visual reasoning. VCTP can adaptively focus on the related visual concepts in the image, transform them into natural language, and provide consistent rationales to support the answer prediction. Compared with existing methods, it not only achieves better performance but also maintains high transparency by keeping the whole trace of each reasoning step. We hope that VCTP can motivate future research on interactions between models of different modalities for more effective and interpretable visual commonsense reasoning systems.

Acknowledgements

This work was supported by gift funding from Cisco and Amazon. We thank Zhicheng Zheng for helping us design the pipeline figure.

References

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. *arXiv*.

Andreas, J.; Rohrbach, M.; Darrell, T.; and Klein, D. 2016. Neural module networks. In *CVPR*.

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. Vqa: Visual question answering. In *ICCV*.

Brohan, A.; Chebotar, Y.; Finn, C.; Hausman, K.; Herzog, A.; Ho, D.; Ibarz, J.; Irpan, A.; Jang, E.; Julian, R.; et al. 2022. Do As I Can, Not As I Say: Grounding Language in Robotic Affordances. In *CoRL*.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners.

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

Chen, W.; Hu, H.; Chen, X.; Verga, P.; and Cohen, W. W. 2022a. MuRAG: Multimodal Retrieval-Augmented Generator for Open Question Answering over Images and Text. In *EMNLP*.

Chen, X.; Wang, X.; Changpinyo, S.; Piergiovanni, A.; Padlewski, P.; Salz, D.; Goodman, S.; Grycner, A.; Mustafa, B.; Beyer, L.; et al. 2022b. Pali: A jointly-scaled multilingual language-image model. *arXiv*.

Chen, Z.; Mao, J.; Wu, J.; Wong, K.-Y. K.; Tenenbaum, J. B.; and Gan, C. 2021b. Grounding Physical Concepts of Objects and Events Through Dynamic Visual Reasoning. In *ICLR*.

Chen, Z.; Yi, K.; Li, Y.; Ding, M.; Torralba, A.; Tenenbaum, J.; and Gan, C. 2022c. ComPhy: Compositional Physical Reasoning of Objects and Events from Videos. In *ICLR*.

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

Creswell, A.; Shanahan, M.; and Higgins, I. 2022. Selection-Inference: Exploiting Large Language Models for Interpretable Logical Reasoning. *arXiv*.

Ding, M.; Chen, Z.; Du, T.; Luo, P.; Tenenbaum, J.; and Gan, C. 2021. Dynamic visual reasoning by learning differentiable physics models from video and language. 34: 887–899.

Ding, M.; Xu, Y.; Chen, Z.; Cox, D. D.; Luo, P.; Tenenbaum, J. B.; and Gan, C. 2022a. Embodied Concept Learner: Self-supervised Learning of Concepts and Mapping through Instruction Following. In *CoRL*.

Ding, Y.; Yu, J.; Liu, B.; Hu, Y.; Cui, M.; and Wu, Q. 2022b. MuKEA: Multimodal Knowledge Extraction and Accumulation for Knowledge-based Visual Question Answering. In *CVPR*.

Dou, Z.-Y.; Kamath, A.; Gan, Z.; Zhang, P.; Wang, J.; Li, L.; Liu, Z.; Liu, C.; LeCun, Y.; Peng, N.; et al. 2022. Coarse-to-fine vision-language pre-training with fusion in the backbone. *arXiv*.

Gao, F.; Ping, Q.; Thattai, G.; Reganti, A.; Wu, Y. N.; and Natarajan, P. 2022. Transform-Retrieve-Generate: Natural Language-Centric Outside-Knowledge Visual Question Answering. In *CVPR*.

Gardères, F.; Ziaeefard, M.; Abeloos, B.; and Lecue, F. 2020. Conceptbert: Concept-aware representation for visual question answering. In *Findings of EMNLP*.

Ge, C.; Huang, R.; Xie, M.; Lai, Z.; Song, S.; Li, S.; and Huang, G. 2022. Domain Adaptation via Prompt Learning.

Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *CVPR*.

Gui, L.; Wang, B.; Huang, Q.; Hauptmann, A.; Bisk, Y.; and Gao, J. 2022. KAT: A Knowledge Augmented Transformer for Vision-and-Language. In *NAACL*.

Guo, Y.; Nie, L.; Wong, Y.; Liu, Y.; Cheng, Z.; and Kankanhalli, M. 2022. A Unified End-to-End Retriever-Reader Framework for Knowledge-based VQA. In *ACM MM*.

Han, X.; Yang, J.; Hu, H.; Zhang, L.; Gao, J.; and Zhang, P. 2021. Image scene graph generation (sgg) benchmark. *arXiv*.

Hudson, D. A.; and Manning, C. D. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 6700–6709.

Jia, M.; Tang, L.; Chen, B.-C.; Cardie, C.; Belongie, S.; Hariharan, B.; and Lim, S.-N. 2022. Visual prompt tuning. In *ECCV*.

Jin, W.; Cheng, Y.; Shen, Y.; Chen, W.; and Ren, X. 2022. A Good Prompt Is Worth Millions of Parameters: Low-resource Prompt-based Learning for Vision-Language Models. In *ACL*.

Johnson, J.; Hariharan, B.; Van Der Maaten, L.; Fei-Fei, L.; Lawrence Zitnick, C.; and Girshick, R. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*.

Ju, C.; Han, T.; Zheng, K.; Zhang, Y.; and Xie, W. 2021. Prompting visual-language models for efficient video understanding. *arXiv*.

Kamath, A.; Clark, C.; Gupta, T.; Kolve, E.; Hoiem, D.; and Kembhavi, A. 2022. Webly Supervised Concept Expansion for General Purpose Vision Models. In *ECCV*.

Konstantopoulos, S. 2010. An Embodied Dialogue System with Personality and Emotions. In *Proceedings of the 2010 Workshop on Companionable Dialogue Systems*, 31–36. Uppsala, Sweden: Association for Computational Linguistics.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv*.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *ICML*.

Lin, W.; and Byrne, B. 2022. Retrieval Augmented Visual Question Answering with Outside Knowledge. In *EMNLP*.

Lu, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*.

Lu, J.; Clark, C.; Zellers, R.; Mottaghi, R.; and Kembhavi, A. 2022. Unified-io: A unified model for vision, language, and multi-modal tasks. *arXiv*.

Marasovic, A.; Beltagy, I.; Downey, D.; and Peters, M. 2022. Few-Shot Self-Rationalization with Natural Language Prompts. In *Findings of NAACL*.

Marino, K.; Chen, X.; Parikh, D.; Gupta, A.; and Rohrbach, M. 2021. Krisp: Integrating implicit and symbolic knowledge for open-domain knowledge-based vqa. In *CVPR*.

Marino, K.; Rastegari, M.; Farhadi, A.; and Mottaghi, R. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *CVPR*.

Oaksford, M.; and Chater, N. 2007. *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford University Press.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *NIPS*.

Schwenk, D.; Khandelwal, A.; Clark, C.; Marino, K.; and Mottaghi, R. 2022. A-OKVQA: A Benchmark for Visual Question Answering using World Knowledge.

Stenning, K.; and Van Lambalgen, M. 2012. *Human reasoning and cognitive science*. MIT Press.

Tan, H.; and Bansal, M. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In *EMNLP*.

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

Tsimpoukelli, M.; Menick, J. L.; Cabi, S.; Eslami, S.; Vinyals, O.; and Hill, F. 2021. Multimodal few-shot learning with frozen language models. In *NeurIPS*.

Wang, P.; Wu, Q.; Shen, C.; Dick, A.; and Van Den Hengel, A. 2017a. Fvqa: Fact-based visual question answering. *PAMI*.

Wang, P.; Wu, Q.; Shen, C.; Dick, A. R.; and van den Hengel, A. 2017b. Explicit Knowledge-based Reasoning for Visual Question Answering. In *IJCAI*.

Wang, Z.; Li, M.; Xu, R.; Zhou, L.; Lei, J.; Lin, X.; Wang, S.; Yang, Z.; Zhu, C.; Hoiem, D.; et al. 2022. Language Models with Image Descriptors are Strong Few-Shot Video-Language Learners. *arXiv*.

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

Wu, J.; Lu, J.; Sabharwal, A.; and Mottaghi, R. 2022. Multimodal answer validation for knowledge-based vqa. In *AAAI*.

Yang, Z.; Gan, Z.; Wang, J.; Hu, X.; Lu, Y.; Liu, Z.; and Wang, L. 2022. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *AAAI*.

Yi, K.; Gan, C.; Li, Y.; Kohli, P.; Wu, J.; Torralba, A.; and Tenenbaum, J. B. 2020. Clevrer: Collision events for video representation and reasoning. In *ICLR*.

Yi, K.; Wu, J.; Gan, C.; Torralba, A.; Kohli, P.; and Tenenbaum, J. B. 2018. Neural-Symbolic VQA: Disentangling Reasoning from Vision and Language Understanding. In *NIPS*.

Yu Jiang*; Vivek Natarajan*; Xinlei Chen*; Rohrbach, M.; Batra, D.; and Parikh, D. 2018. Pythia v0.1: the Winning Entry to the VQA Challenge 2018. *arXiv*.

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

Zeng, A.; Attarian, M.; Ichter, B.; Choromanski, K.; Wong, A.; Welker, S.; Tombari, F.; Purohit, A.; Ryoo, M.; Sindhwani, V.; Lee, J.; Vanhoucke, V.; and Florence, P. 2022. Socratic Models: Composing Zero-Shot Multimodal Reasoning with Language. *arXiv*.

Zhang, S.; Roller, S.; Goyal, N.; Artetxe, M.; Chen, M.; Chen, S.; Dewan, C.; Diab, M.; Li, X.; Lin, X. V.; et al. 2022. Opt: Open pre-trained transformer language models. *arXiv*.

Zhang, Z.; Zhang, A.; Li, M.; Zhao, H.; Karypis, G.; and Smola, A. J. 2023. Multimodal Chain-of-Thought Reasoning in Language Models. *ArXiv*, abs/2302.00923.

Zhao, C.; Shen, Y.; Chen, Z.; Ding, M.; and Gan, C. 2023. TextPSG: Panoptic Scene Graph Generation from Textual Descriptions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2839–2850.

Zhong, Y.; Yang, J.; Zhang, P.; Li, C.; Codella, N.; Li, L. H.; Zhou, L.; Dai, X.; Yuan, L.; Li, Y.; and Gao, J. 2022. RegionCLIP: Region-Based Language-Image Pretraining. In *CVPR*.

Zhou, D.; Schärli, N.; Hou, L.; Wei, J.; Scales, N.; Wang, X.; Schuurmans, D.; Bousquet, O.; Le, Q.; and Chi, E. 2022a. Least-to-Most Prompting Enables Complex Reasoning in Large Language Models. *arXiv*.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022b. Conditional Prompt Learning for Vision-Language Models. In *CVPR*.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022c. Learning to prompt for vision-language models. *IJCV*.

Zhu, Z.; Yu, J.; Wang, Y.; Sun, Y.; Hu, Y.; and Wu, Q. 2021. Mucko: multi-layer cross-modal knowledge reasoning for fact-based visual question answering. In *IJCAI*.