Can Frozen Large Language Models Solve Visual Reasoning?

Anonymous ACL submission

Abstract

We present ReasonLM, a simple framework which utilizes a pre-trained, frozen large language model (LLM) for visual reasoning tasks, and achieves competitive performance on ACRE and MEWL. We demonstrate for the first time that a frozen LLM serves as a task-agnostic reasoning machine for diverse reasoning tasks that involve object recognition, causal induction, and relation modeling. ReasonLM does not rely on synthesizing symbolic programs or self-supervised visual representation learning. Rather, it learns an object-centric, light-weight visual encoder from scratch. Via its simplified design, we investigate the essential design choices for strong visual reasoning performance. Code and model will be released.

1 Introduction

Visual reasoning tasks examine the abilities of extracting visual information, recognizing the relations and patterns among the visual information, and generalizing to novel situations by making analogies (Zhang et al., 2021; Jiang et al., 2023; Zhang et al., 2019; Chollet, 2019; Moskvichev et al., 2023; Girdhar and Ramanan, 2019). Thus, such tasks evaluate a model's visual perception capabilities and logical reasoning capabilities, both of which reflect how intelligent a model is.

There has been a series of work studying various approaches to solve visual reasoning tasks. Some approaches rely on task-specific visual encoders, such as symbolic object encoders (Zhang et al., 2021), object detectors (Ding et al., 2021), or on task-specific training strategies for these visual encoders (Sun et al., 2024; Bhattacharyya et al., 2023). Other approaches introduce inductive biases by developing task-specific visual reasoning modules (Hu et al., 2021; Benny et al., 2021). However, these task-specific components limit the scalability and generalizability across different visual reasoning tasks.

In this work, we investigate how to simplify a visual reasoning framework in order to minimize the task-specific designs and maximize the sharing of visual encoders and reasoning modules across tasks. We propose a visual reasoning framework ReasonLM, which consists of a perception module and a task-agnostic reasoning module. The perception module is a light-weight visual encoder which does not require large-scale self-supervised pretraining or task-specific inductive biases, such as slot attention layers (Sun et al., 2024) and interleaved cross-attention mechanism (Bhattacharyya et al., 2023). The reasoning module is a frozen pretrained large language model (LLM) which solves visual reasoning tasks as sequence modeling problems, and simplifies the pretraining of the visual encoder by optimizing the visual encoder with a next token prediction objective.

We focus on two visual reasoning tasks, ACRE (Zhang et al., 2021) and MEWL (Jiang et al., 2023), which require a model to recognize and infer the implicit or explicit properties of objects with a limited number of observations and generalize to new scenarios. We first show that a frozen pretrained LLM can be used as the underlying reasoner and shared across different visual reasoning tasks, when these tasks represent the images into symbolic representations. Following, we study if frozen LLMs can solve visual reasoning with image inputs by projecting image representations into language latent space. We demonstrate that, for the first time, state-of-the-art visual reasoning can be achieved by learning a simple 2-layer ViT encoder from scratch. Last, we investigate if a learned visual encoder combined with a frozen LLM can solve a visual reasoning task which requires understanding of a specific concept, to what extent this model can solve other tasks involving the same concept (Moskvichev et al., 2023). We observe that the visual encoders can be adapted to other reasoning sub-tasks with a simply learned linear projection.

At last, we identify two essential design choices for the visual encoder: (1) Object-centric inputs or representations are crucial for solving visual reasoning tasks; (2) Visual encoders learned with a diverse set of sub-tasks are easier to transfer to novel tasks.

2 Related Works

While deep neural networks have achieved considerable success on image understanding tasks that require reasoning, such as visual question answering (Hudson and Manning, 2019b; Zellers et al., 2019; Marino et al., 2019), benchmarks collected "from-the-wild" often contain dataset and language biases, which make it harder to rigorously measure progress (Goyal et al., 2017). As a response, a series of synthetic, diagnostic datasets (Johnson et al., 2017; Yi et al., 2019; Girdhar and Ramanan, 2019; Zhang et al., 2021) have been proposed to benchmark visual reasoning. We investigate the use of a frozen large language model in visual reasoning tasks. LLMs have been previously applied to reasoning with natural language (Magister et al., 2022; Wei et al., 2022; Chen et al., 2022; Liu et al., 2020), but the scope of "reasoning" they can rigorously solve is questioned (Kambhampati, 2024; Mitchell et al., 2023; Gendron et al., 2024). Our paper aims to leverage LLM as a tool to assist visual reasoning, and our problem scope is defined by the ACRE (Zhang et al., 2021) and MEWL (Jiang et al., 2023) benchmarks, which provide a training dataset of example context observations, query observation, and the desirable output to illustrate the reasoning tasks of interest.

Recent approaches on visual reasoning can be categorized into neuro-symbolic methods (Mao et al., 2019; Hudson and Manning, 2019a), or neural networks with implicit representations (Ding et al., 2021; Sun et al., 2024; Bhattacharyya et al., 2023). Both approaches roughly follow the same outline of perception stage and reasoning stage. The outputs of the perception stage are usually object-centric, which can be obtained with a supervised object detector (Traub et al., 2023; He et al., 2017), or with self-supervised object discovery (Geirhos et al., 2018; Hermann et al., 2020; Olah et al., 2017; Burgess et al., 2019; Locatello et al., 2020; Caron et al., 2021). It has been observed that on diagostic datasets, both approaches lead to satisfactory object localization. For reasoning, the former approach generates an interpretable program to execute on the recognized visual inputs, while the latter approach often relies on a reasoning neural network trained specifically to a certain reasoning task. Our paper investigates the use of a frozen LLM as the shared reasoning module, where the visual encoders are trained from scratch.

3 Visual Reasoning Tasks

In this work, we consider visual reasoning tasks as a task which requires forming and abstracting concepts, and making generalization to new problems ,via analogies, from a limited number of observations. We focus on visual reasoning tasks that are based on CLEVR objects (Johnson et al., 2017) such that we can obtain oracle visual perception.

For each problem in a visual reasoning task, there are N pairs of context frame and label $\{c_i, l_c^i\}_{i=1}^N$ and a query frame q. On each frame, there are objects which can be represented by object attributes (i.e., color, material, shape) and their location information (i.e., bounding boxes). The task is to solve the query frame by inducing the patterns in context frames and applying the patterns on the query frame.

ACRE (Zhang et al., 2021) evaluates a model's ability of causal induction, which means to identify unobservable causal relationships from limited number of observations. It is inspired by Blicket detection experiments from developmental psychology (Gopnik and Sobel, 2000), where a Blicket detector will be activated when at least one Blicket object is placed on it. Since the Blicket-ness is an unobservable property of the objects, it is needed to infer which objects are Blickets by observing several context trials where different combinations of objects are placed on Blicket detector, revealing its activation status. In ACRE, there are 6 context frames and 4 query frames per sample, where each frame contains a distinct set of CLEVR objects (Johnson et al., 2017). Given the context frames and a query frame, a model needs to predict the activation status of Blicket detector in query frame, which can be *on*, *off*, or *unknown*. We mask out the Blicket detectors in all the frames in order to avoid our model to directly infer activation status by looking at the Blicket detectors.

MEWL (Jiang et al., 2023) evaluates a model's ability of novel word learning in grounded visual scenes. It simulates children's word learning process which is inherently few-shot and open-ended and contains referential uncertainty. In MEWL,

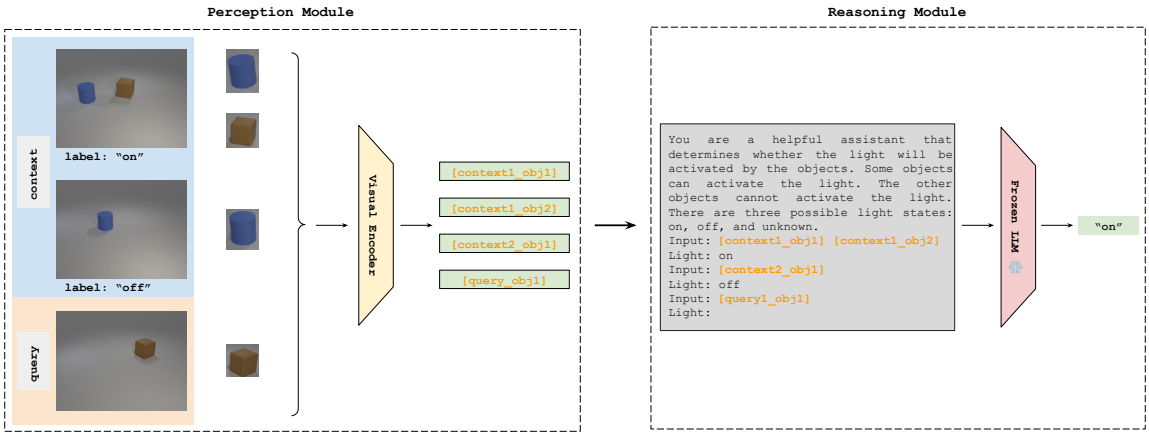


Figure 1: Overflow of ReasonLM, specifically a ReasonLM-Object. ReasonLM consists of a perception module and a reasoning module. Perception module is a visual encoder which takes object crops as inputs and outputs object representations. The object representations are passed into task-specific prompts to represent panels. Reasoning module is a frozen LLM which consumes the task-specific prompts and predicts the answer for the query panel.

there are 9 different sub-tasks: *shape*, *color*, *material*, *object*, *composite*, *relation*, *bootstrap*, *number*, *pragmatic*. These 9 tasks cover four types of scenarios: basic attribute naming, relational word learning, number word learning, and pragmatic word learning. For each data sample, there are 6 context frames and 1 query frame per sample, where each frame contains a set of CLEVR objects and a corresponding novel word or phrase (i.e., *utterance*). The task is to understand the meaning of the novel words by observing the context frames, and select the correct utterances out of 5 options for the query frame.

4 Can Frozen LLMs Help Reasoning?

Before studying whether a frozen pretrained LLM can solve visual reasoning, we first investigate whether an LLM can serve as a reasoning module. Thus, we assume oracle perception is available, and evaluate how well a frozen pretrained LLM can solve ACRE and MEWL with oracle information.

4.1 Prompts for Reasoning Tasks

We follow the prompt design in Gendron et al. (2024) such that there are two parts for each prompt. First, a task definition is used to define the visual reasoning task. Second, the descriptions of each panel and its corresponding label. On ACRE and MEWL, we retrieve panel captions which contain oracle information of the panels. We then directly use these captions to represent each panel, and obtain the prompts for ACRE and MEWL. On ACRE, each panel caption describes all the objects appeared on a panel (e.g., "There are brown rubber sphere and cyan metal cylinder."). On MEWL, each panel caption is generated by the panel captioner from Jiang et al. (2023).

4.2 Language Baseline

A language baseline is where a frozen LLM directly consumes visual reasoning prompts and make predictions by selecting the answer option with the hightest joint probability. This setup is aligned with the multiple choice evaluation in Gendron et al. (2024).

4.3 ReasonLM

We introduce ReasonLM framework which leverages a frozen pretrained LLM as the reasoning module (Figure 1). The inputs to the LLM are task-specific prompts. The difference between ReasonLM and the language baseline is how a panel is represented. For ReasonLM, a panel is represented by object representations. For each object in a panel, a symbolic encoder[1] encodes object attributes with corresponding embedding layers and encodes objects' location information[2] with a linear layer. Following, the attribute embeddings and location representation are concatenated and passed into a projection layer to obtain an object representation in token embedding space. These object representations can be considered as projected object tokens and are passed to the prompts to represent

[1] As opposed to updating the token embedding directly, we select the symbolic encoder which decouples the input representations with general task definitions in the prompts.

[2] Each object location is represented as $[x_1, y_1, x_2, y_2, w, h, w \times h]$

panels. LLM will take in the prompts with these object tokens and make predictions based on the highest joint probabilities of the answer options.

The symbolic encoder is randomly initialized and trained from scratch. During training, the pretrained LLM will be frozen, and the symbolic encoder is trained with next token prediction objective. Following, we refer this ReasonLM with symbolic encoder as ReasonLM-Symbol.

4.4 Implementation Details

On ACRE, we use the training set which involves 6000 samples, where each sample contains 6 context frames and 4 query frames. Thus, the training set has 24000 sequences. On MEWL, we use the training sets of the 9 sub-tasks, each of which involves 600 samples. Thus, the number of training sequences is 5400. On both datasets, the image frames are resized to 224×224 and will be tokenized to image patches with size of 16×16.

We choose the pretrained LLaMA2 (Touvron et al., 2023) with 7 billion parameters as our reasoning module. We use a symbolic encoder with the embedding size of 32 for each object property.

During pretraining, we use the AdamW optimizer with a learning rate of 3×10^{-5}. We pretrain the visual encoders for 20 epochs on ACRE, and 40 epochs on MEWL. The batch size is set to 64.

Method	Projection	ACRE(%)	MEWL(%)
Random	-	33.3	20.0
LLaMA2-7B	-	39.9	39.2
GPT-2*	-	37.1	-
GPT-3.5-Turbo*	-	18.4	-
GPT-4*	-	27.2	-
Alpaca*	-	3.6	-
ReasonLM-Symbol	Linear	92.0	69.1
ReasonLM-Symbol	MLP	98.5	-

Table 1: Results on visual reasoning tasks with oracle perception. Our ReasonLM-Symbol with linear projection significantly outperforms language baseline (LLaMA2-7B) on both ACRE and MEWL, indicating that LLaMA-7B has sufficient abstract reasoning capability to solve visual reasoning tasks if oracle visual perception is provided as inputs. Results with * are language baselines from (Gendron et al., 2024).

4.5 Results

Results are shown in Table 1. We observe that ReasonLM-Symbol can significantly outperform its language baseline on both ACRE and MEWL by simply learning the simple symbolic encoder without updating the weights of the LLM backbone. This indicates that a frozen pretrained LLM can serve as the reasoning module for visual reasoning tasks on ACRE and MEWL when oracle perception is available. We further experiment with a ReasonLM-Symbol with a 2-layer MLP as the final projection layer on ACRE. We observe that the increase in complexity of the final projection layer can further improve model's performance to nearly perfect on ACRE. This further supports that LLMs can reason, but LLMs do not perform the best with natural language inputs. Instead, LLMs may benefit more from linearly or non-linearly encoded information.

5 What Makes Good Visual Representations for LLM Reasoners?

Given that LLMs can be used as the reasoning modules in visual reasoning task when oracle perception is available, we study the factors of useful visual representations for LLMs in order to unlock LLMs' reasoning capabilities with visual inputs in visual reasoning tasks. We mainly consider these factors: (1) Is visual pretraining needed for visual reasoning on ACRE and MEWL? (2) Is object-centric inductive bias needed? (3) Do different model inductive biases make any difference?

5.1 ReasonLM with Image Inputs

ReasonLM-Object takes object crops in each panel as inputs and use a visual encoder to retrieve representations of the objects in each panel. Thus, each panel is represented by a number of object representations. This variant assumes ground truth object detection exists in order to control the factors of reasoning performance. In fact, it is reasonable to learn a good object detector on ACRE and MEWL (Ding et al., 2021). This variant has most of the information needed for solving ACRE and MEWL, except spatial information, since the object location information is missing.

ReasonLM-Image takes each panel as inputs and use a visual encoder to retrieve panel representations. This variant simplifies the inputs the most, but requires the visual encoder to understand object properties and spatial relationships between objects directly from panel images.

Method	shape	color	material	object	composite	relation	bootstrap	num.	pragmatic	Avg.
BERT	94.8	98.8	97.5	19.5	97.8	22.2	62.2	21.8	99.8	68.3
GPT-3.5	96.8	82.3	87.0	98.2	88.3	20.0	45.8	22.7	26.7	63.1
ALOE	34.2	33.2	31.0	19.5	30.5	21.5	27.5	23.3	20.8	26.8
Flamingo-1.1B	49.3	35.3	48.5	19.2	38.2	18.8	57.3	84.2	18.0	41.0
ReasonLM-Symbol	80.8	84.2	82.8	97.7	65.7	18.2	84.8	87.7	19.7	69.1
ReasonLM-Image	31.8	98.8	76.2	24.0	30.7	17.7	36.0	48.7	18.7	42.5
ReasonLM-Object	34.8	99.5	99.5	96.8	53.3	19.7	84.8	99.8	21.3	67.7

Table 2: Results of ReasonLM with image inputs on MEWL tasks. We observe that, though ReasonLM-Object does not require task-specific inductive biases or training strategies, ReasonLM-Object significantly outperforms previous state-of-the-art, indicating that a frozen LLaMA2-7B can perform visual reasoning with image inputs.

Method	Encoder	Acc.(%)
NS-OPT	-	66.3
IV-CL	-	93.0
ReasonLM-Symbol	Linear	92.0
ReasonLM-Image	ViT-L2H4	76.6
ReasonLM-Object	ViT-L2H4	97.0
ReasonLM-Object	ViT-L6H8	96.6
ReasonLM-Object	ResNet-50	93.6

Table 3: Results of ReasonLM with image inputs on ACRE I.I.D. split. ReasonLM-Object significantly outperforms prior works which use self-supervised visual encoders, indicating that visual pretraining is not necessary for visual reasoning tasks. By running an ablation on visual encoder, we observe that model inductive bias does not make a big difference on ACRE.

5.2 Implementation Details

For the visual encoders for ReasonLM-Object and ReasonLM-Image, we use a 2-layer ViT (Dosovitskiy et al., 2020) with 4 attention heads and a 768-dimensional hidden space, and stack a linear projection layer on top at map the hidden representations to token embedding space. The visual encoders are randomly initialized and trained with next token prediction objective with a frozen LLM. The training data and other hyperparameters are the same as ReasonLM-Symbol. In contrast to IV-CL (Sun et al., 2024), where the heavy visual encoder contains model inductive bias (i.e., slot attention (Locatello et al., 2020)), and it needs to be pretrained, ReasonLM-Object and ReasonLM-Image do not require any task-specific pretraining or model inductive biases.

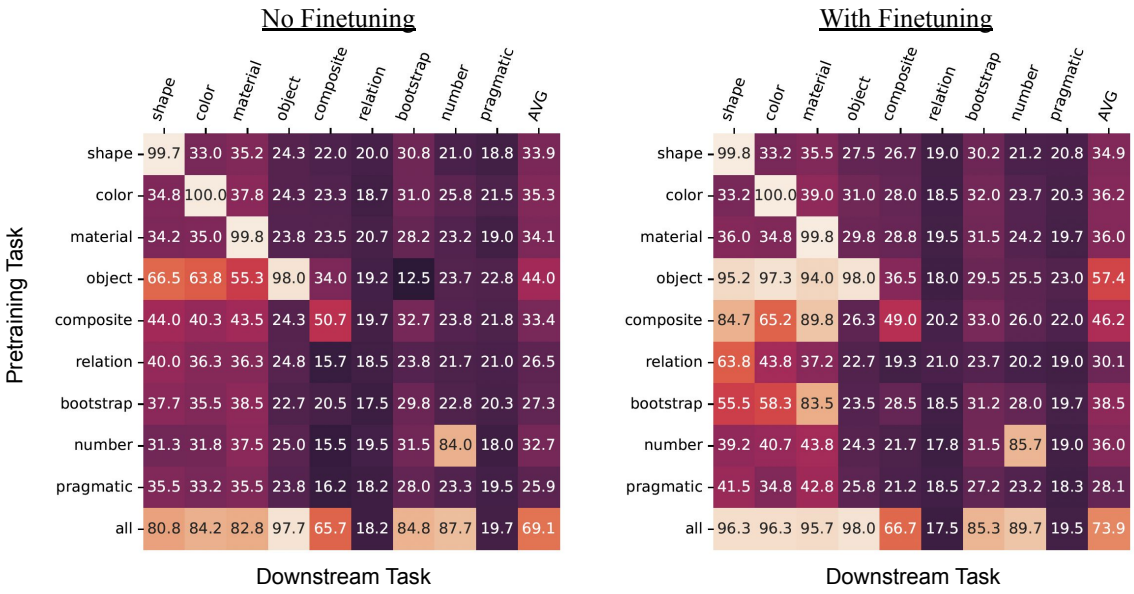
5.3 Results

Tables 2 and 3 summarize the results on ReasonLM with image inputs. We show that on both ACRE and MEWL, ReasonLM-Object achieves strong visual reasoning performance, which is on par or even better than ReasonLM-Symbol which uses oracle symbolic object information. Furthermore, we demonstrate that on ACRE, ReasonLM-Object with a simple 2-layer ViT trained from scratch significantly outperforms IV-CL (Sun et al., 2024) and ALOE (Ding et al., 2021) which use self-supervised visual encoders, indicating that visual pretraining is not necessary for visual reasoning tasks. By comparing ReasonLM-Object and ReasonLM-Image on ACRE and MEWL, we find that object-centric representations are essential for the great performance for ReasonLM-Object, reflecting that input inductive bias is still needed. Last, we run an ablation study of model inductive bias on ACRE. We observe that ReasonLM-Object with either ViT (Dosovitskiy et al., 2020) or ResNet (He et al., 2016) is competitive, though ViT performs slightly better. This indicates that model inductive bias does not make a big difference on ACRE.

6 Can LLM-supervised Visual Representations Generalize Across Different Visual Reasoning Tasks?

With object-centric inductive bias, our ReasonLM-Object performs well on ACRE and MEWL. However, it is unclear whether the LLM-supervised visual encoders are task-specific or are generalizable across different reasoning tasks. In this section, we first study whether the learned visual representations are generalizable across different visual reasoning tasks. Following, we explore how to better transfer the learned visual encoders by finetuning the final linear projection in the visual encoder. At last, we investigate what types of reasoning tasks lead to more generalizable visual representations.

(a) Transfer Learning with ReasonLM-Symbol



(b) Transfer Learning with ReasonLM-Object

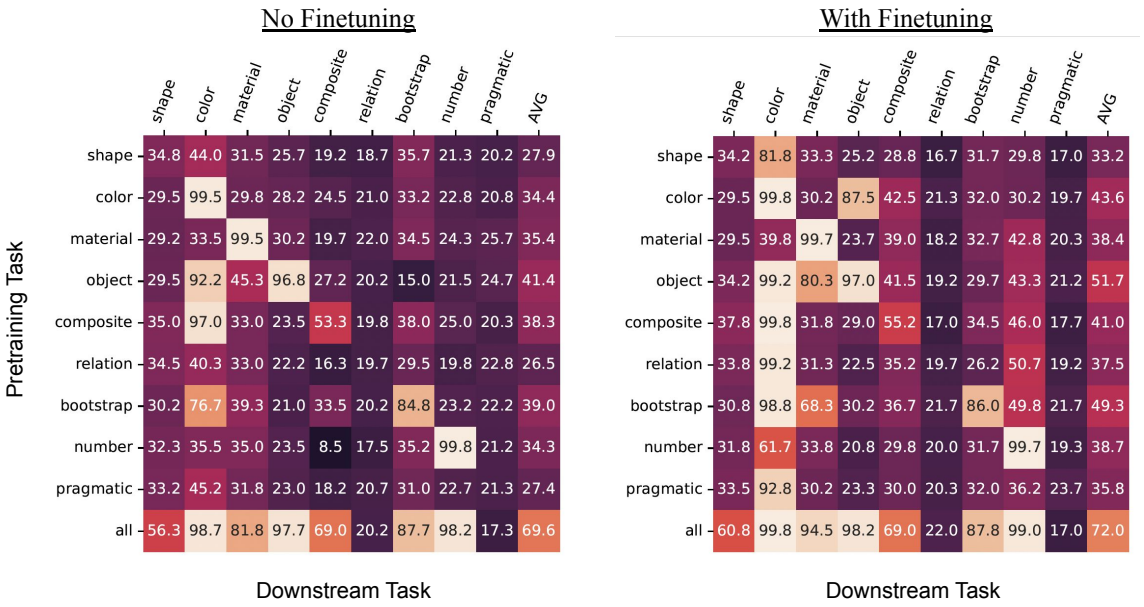


Figure 2: Results of transfer learning experiment on MEWL. For both ReasonLM-Symbol and ReasonLM-Object, the models pretrained on a task perform worse when they are transferred to other tasks, compared to the models pretrained just for this task. This reflect that learned visual encoders are task-specific, and do not generalize directly. After we finetune the final linear projection of the visual encoders, the visual encoders consistently generalize better on all sub-tasks.

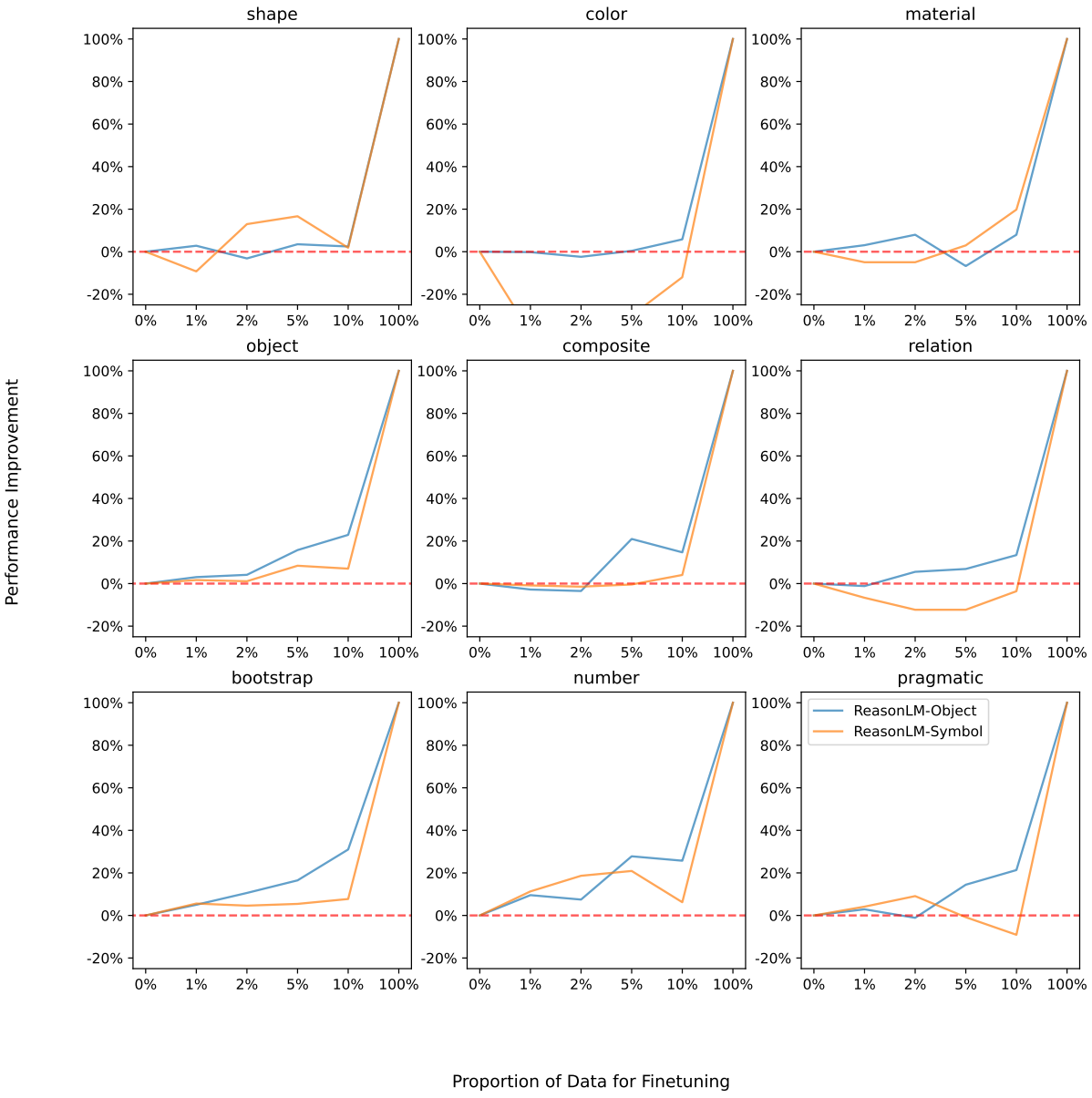


Figure 3: Data efficiency analysis on ReasonLM-Symbol and ReasonLM on MEWL. X-axis is the proportion of data used for finetuning. Y-axis is the amount of performance improvement normalized by min-max normalization. We observe that when the amount of data used to finetune the pretrained visual encoders is low, the finetuned visual encoders do not show strong generalization for both ReasonLM-Symbol and ReasonLM-Object.

6.1 Transfer Learning Experiment

We conduct transfer learning experiment to measure the generalizability of the learned visual representations. Given that there are nine different sub-tasks in MEWL, we focus on MEWL for transfer learning experiment. We consider sub-task training and all-task training. Sub-task training means to pretrain a visual encoder on one sub-task in MEWL, and all-task training means to pretrain a visual encoder on all the nine sub-tasks together. After we train ReasonLM-Object on each sub-task and on all-task, we conduct transfer learning experiment by applying a ReasonLM-Object pretrained on a sub-task A to another sub-task B. To further explore how to better transfer the learned visual encoders, we finetune only the last linear projection in a visual encoder during the transfer learning experiment.

6.2 Results

Results are shown in Figure 2. For both ReasonLM-Symbol and ReasonLM-Object, the models pretrained on a task perform worse when they are transferred to other tasks, compared to the models pretrained just for this task. The only special case is sub-task `color`, where ReasonLM-Object pretrained on sub-tasks `object`, `composite` and `bootstrap` can perform reasonably well on sub-task `color`. We attribute this to the fact that color is one of the basic object property which is required to solve the problems in these sub-tasks. In all, these results reflect that the learned visual encoders are task-specific, and do not generalize directly.

Next, we observe that with the finetuned final linear projection, the visual encoders consistently generalize better on all sub-tasks. We find that reasoning tasks do make a difference on the generalizability of the learned encoders. When a task requires the understanding of more visual information, the better the visual encoders pretrained on it can transfer to other tasks. For example, among all sub-tasks, pretraining with `object` sub-task works the best for both ReasonLM-Symbol and ReasonLM-Object.

7 Data Efficiency of LLM-supervised Visual Encoders for Transfer Learning

If a model can performs abstract reasoning, then it is expected to see this model to generalize across different tasks with few-shot examples (Chollet, 2019; Moskvichev et al., 2023; Mitchell, 2021). Therefore, we explore how data efficient our pretrained visual encoders are for transfer learning.

7.1 Experimental Setup

We still focus on MEWL and all the setup remains the same as in transfer learning experiment in Section 6, except that we only use 1%/2%/5%/10% of the training data from a sub-task A to finetune the last linear projection of a visual encoder pretrained on sub-task B. After a pretrained visual encoder is finetuned on a new task, we evaluate how well this visual encoder can perform on this new sub-task.

7.2 Results

Figure 3 shows the results, where we carry out a min-max normalization on the amount of performance improvement based on the performance without finetuning and performance with full finetuning. For example, for ReasonLM-Symbol's visual encoder pretrained on `shape` sub-task, if we finetune it with 5% data, we can reach about 20% of the performance improvement, compared to full finetuning. We observe that when the amount of data used to finetune the pretrained visual encoders is low, the finetuned visual encoders do not show strong generalization for both ReasonLM-Symbol and ReasonLM-Object. This indicates that we are still far away from performing abstract reasoning, and more work needs to be done to investigate whether the pretrained visual encoders perform parts of the reasoning job and how to make this visual reasoning system more data efficient.

8 Conclusions

We present a simple yet effective framework for visual reasoning, powered by a pre-trained, frozen large language model. We demonstrate that LLMs can solve visual reasoning given both oracle object information or image observations. Unlike previous approaches that rely on task-specific reasoning modules or pre-trained visual encoders, our proposed ReasonLM uses the frozen LLM as the shared reasoning module, and works by training a light-weight visual encoder from scratch. Through evaluations on ACRE and MEWL, we demonstrate that ReasonLM achieves competitive reasoning performance, as long as the visual representations are object-centric. Moreover, we demonstrate that the visual encoders can be transferred across different reasoning tasks, subject to a linear projection.

9 Limitations

We conduct evaluations on synthetic datasets only. While the reasoning setup is realistic and designed to avoid dataset bias, recognizing objects and their attributes is arguably much simpler than from images captured in the real world. Additionally, the visual encoders, though light-weight and can be trained from scratch, are specific to individual reasoning (sub-)tasks. Learning generalizable or easily transferable visual representations remain an important open problem.

10 Ethics Statement

We study the use of a pre-trained large language model for visual reasoning tasks. All benchmarks we used for evaluation are publicly available. ACRE was released under GPL-3.0 License, and MEWL was released under MIT License.

References

Yaniv Benny, Niv Pekar, and Lior Wolf. 2021. Scale-localized abstract reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12557–12565.

Apratim Bhattacharyya, Sunny Panchal, Reza Pourreza, Mingu Lee, Pulkit Madan, and Roland Memisevic. 2023. Look, remember and reason: Grounded reasoning in videos with language models. In *The Twelfth International Conference on Learning Representations*.

Christopher P Burgess, Loic Matthey, Nicholas Watters, Rishabh Kabra, Irina Higgins, Matt Botvinick, and Alexander Lerchner. 2019. Monet: Unsupervised scene decomposition and representation. *arXiv preprint arXiv:1901.11390*.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W Cohen. 2022. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *arXiv preprint arXiv:2211.12588*.

François Chollet. 2019. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*.

David Ding, Felix Hill, Adam Santoro, Malcolm Reynolds, and Matt Botvinick. 2021. Attention over learned object embeddings enables complex visual reasoning. *Advances in neural information processing systems*, 34:9112–9124.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. 2018. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*.

Gaël Gendron, Qiming Bao, Michael Witbrock, and Gillian Dobbie. 2024. Large language models are not strong abstract reasoners. *arXiv preprint arXiv:2305.19555*.

Rohit Girdhar and Deva Ramanan. 2019. Cater: A diagnostic dataset for compositional actions and temporal reasoning. *arXiv preprint arXiv:1910.04744*.

Alison Gopnik and David M Sobel. 2000. Detecting blickets: How young children use information about novel causal powers in categorization and induction. *Child development*, 71(5):1205–1222.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Katherine Hermann, Ting Chen, and Simon Kornblith. 2020. The origins and prevalence of texture bias in convolutional neural networks. *Advances in Neural Information Processing Systems*, 33:19000–19015.

Sheng Hu, Yuqing Ma, Xianglong Liu, Yanlu Wei, and Shihao Bai. 2021. Stratified rule-aware network for abstract visual reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1567–1574.

Drew Hudson and Christopher D Manning. 2019a. Learning by abstraction: The neural state machine. *Advances in Neural Information Processing Systems*, 32.

Drew A Hudson and Christopher D Manning. 2019b. Gqa: A new dataset for real-world visual reasoning

and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Guangyuan Jiang, Manjie Xu, Shiji Xin, Wei Liang, Yujia Peng, Chi Zhang, and Yixin Zhu. 2023. Mewl: Few-shot multimodal word learning with referential uncertainty. In *International Conference on Machine Learning*, pages 15144–15169. PMLR.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910.

Subbarao Kambhampati. 2024. Can large language models reason and plan? *Annals of the New York Academy of Sciences*, 1534(1):15–18.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2020. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. *arXiv preprint arXiv:2007.08124*.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. 2020. Object-centric learning with slot attention. *Advances in Neural Information Processing Systems*, 33:11525–11538.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. 2022. Teaching small language models to reason. *arXiv preprint arXiv:2212.08410*.

Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B Tenenbaum, and Jiajun Wu. 2019. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. *arXiv preprint arXiv:1904.12584*.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204.

Melanie Mitchell. 2021. Abstraction and analogy-making in artificial intelligence. *Annals of the New York Academy of Sciences*, 1505(1):79–101.

Melanie Mitchell, Alessandro B Palmarini, and Arseny Moskvichev. 2023. Comparing humans, gpt-4, and gpt-4v on abstraction and reasoning tasks. *arXiv preprint arXiv:2311.09247*.

Arseny Moskvichev, Victor Vikram Odouard, and Melanie Mitchell. 2023. The conceptarc benchmark: Evaluating understanding and generalization in the arc domain. *arXiv preprint arXiv:2305.07141*.

Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. 2017. Feature visualization. *Distill*. Https://distill.pub/2017/feature-visualization.

Chen Sun, Calvin Luo, Xingyi Zhou, Anurag Arnab, and Cordelia Schmid. 2024. Does visual pretraining help end-to-end reasoning? *Advances in Neural Information Processing Systems*, 36.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Manuel Traub, Sebastian Otte, Tobias Menge, Matthias Karlbauer, Jannik Thuemmel, and Martin V Butz. 2023. Learning what and where: Disentangling location and identity tracking without supervision. In *The Eleventh International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

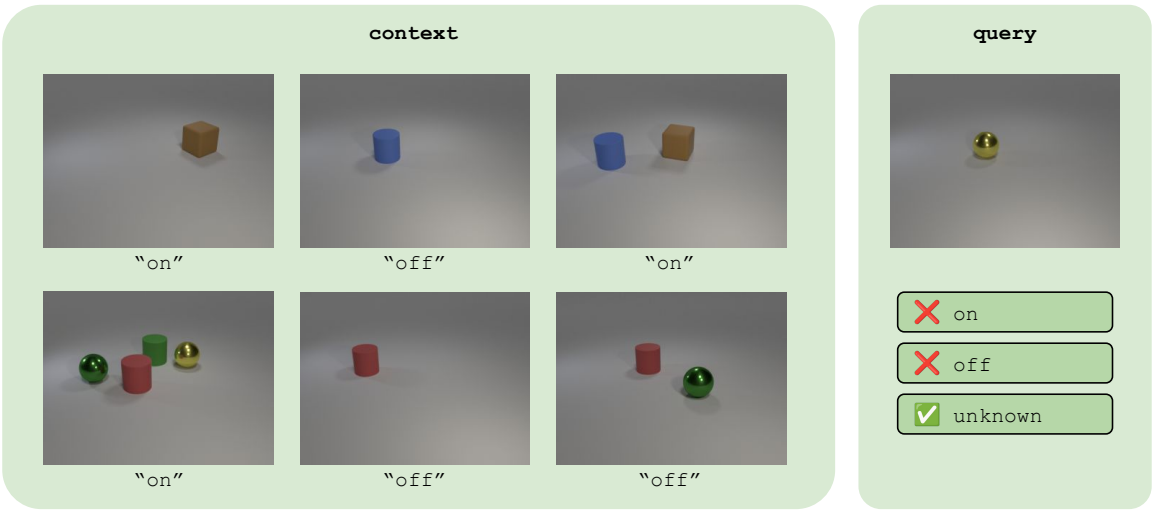
Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. 2019. Clevrer: Collision events for video representation and reasoning. *arXiv preprint arXiv:1910.01442*.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6720–6731.

Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. 2019. Raven: A dataset for relational and analogical visual reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5317–5327.

Chi Zhang, Baoxiong Jia, Mark Edmonds, Song-Chun Zhu, and Yixin Zhu. 2021. Acre: Abstract causal reasoning beyond covariation. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 10643–10653.

(a) ACRE



(a) MEWL

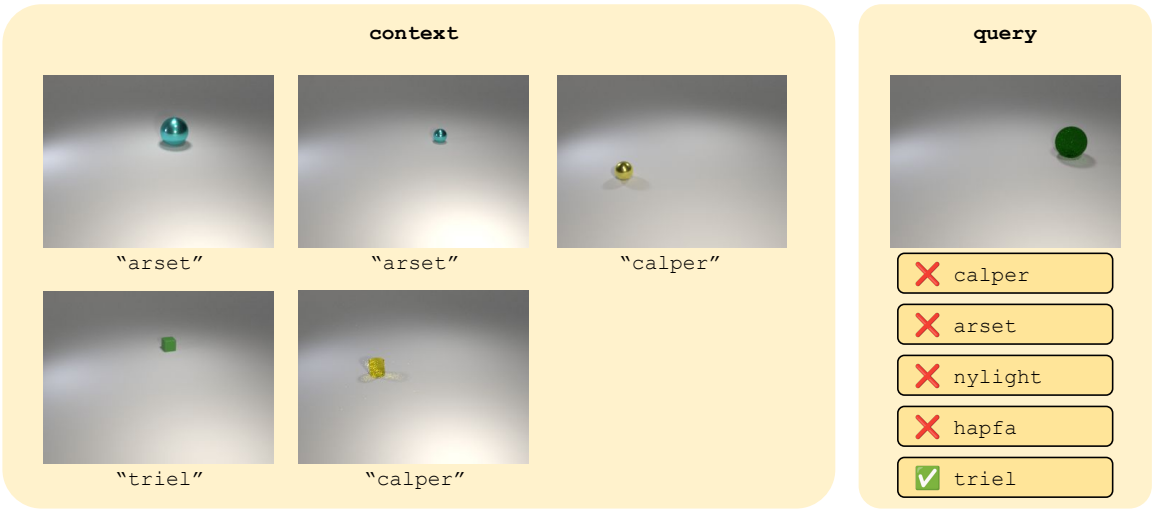


Figure A1: Data samples of ACRE and MEWL. ACRE examines a model's ability of causal induction, which is to identify the unobservable causal relationships from limited number of observations. MEWL tests a model's ability of novel word learning from limited number of observations.