MetaVL: Transferring In-Context Learning Ability From Language Models to Vision-Language Models

Masoud Monajatipoor
UCLA
monajati@ucla.edu

Liunian Harold Li *
UCLA
liunian.harold.li@cs.ucla.edu

Mozhdeh Rouhsedaghat *
USC
rouhseda@usc.edu

Lin F. Yang
UCLA
linyang@ee.ucla.edu

Kai-Wei Chang
UCLA
kwchang@cs.ucla.edu

Abstract

Large-scale language models have shown the ability to adapt to a new task via conditioning on a few demonstrations (i.e., in-context learning). Large-scale language models have shown the ability to adapt to a new task via conditioning on a few demonstrations (i.e., in-context learning). However, in the vision-language domain, most large-scale pre-trained vision-language (VL) models do not possess the ability to conduct in-context learning. How can we enable in-context learning for VL models? In this paper, we study an interesting hypothesis: can we transfer the in-context learning ability from the language domain to the VL domain? Specifically, we first meta-trains a language model to perform in-context learning on NLP tasks (as in MetaICL); then we transfer this model to perform VL tasks by attaching a visual encoder. Our experiments suggest that indeed in-context learning ability can be transferred cross modalities: our model considerably improves the in-context learning capability on VL tasks and can even compensate for the size of the model significantly. On VQA, OK-VQA, and GQA, our method could outperform the baseline model while having ∼20 times fewer parameters.

1 Introduction

Pre-trained language models have shown impressive performance on a range of tasks by learning from large-scale text corpus (Radford et al., 2018, 2019; Yang et al., 2019). Recent studies find that some of these language models can be used to perform *in-context learning* out-of-the-box, i.e., adapting to a task by conditioning on a few demonstrations in context without any gradient update (Brown et al., 2020; Min et al., 2022), which is highly desirable.

In VL modeling, in-context learning is less explored and only a handful of models are proposed to perform in-context learning mainly by limiting the amount of deviation of a pretrained large-scale language model from the language space and translating visual inputs to language embedding space. They either require a large capacity (Tsimpoukelli et al., 2021; Alayrac et al., 2022) or a giant corpus consisting of in-context learning examples (Alayrac et al., 2022; Liu et al., 2023; Koh et al., 2023).

In this work, we explore whether we could enable in-context learning in VL tasks without resorting to extreme scale-up. We study an interesting hypothesis: can we transfer the in-context learning ability from the language domain to the VL domain? To elaborate, not every language model exhibits excellent *in-context* learning ability; recent studies (Min et al., 2022) show that one could explicitly train language models to perform in-context learning, by training the model on multiple tasks with in-context few-shot examples, a process that resembles meta-learning. Thus, an intriguing query arises: when a language model is first meta-trained to perform in-context learning, can it be transferred to perform in-context learning for VL tasks better?

A remarkable observation in our study is the utilization of a meta-trained language model as the transformer encoder-decoder and the mapping of visual features to the language embedding space. This innovative approach led to the development of our proposed VL model (we name it MetaVL). Impressively, our experimental results demonstrate that MetaVL surpasses the baseline model's performance, even when MetaVL is designed to be 20 times smaller in size.

This study makes three main contributions: 1) To the best of our knowledge, this is the first attempt to transfer the meta-learning knowledge for in-context learning from single-modality to multimodality. 2) We propose a VL model, MetaVL[1], which outperforms the baseline in in-context learn-

*equal contribution

[1] https://github.com/masoud-monajati/MetaVL

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 2: Short Papers, pages 495–508

ing while having a much smaller model size. 3) Through extensive experiments on VQA, GQA and OK-VQA, we demonstrate the in-context learning capability of MetaVL and analyze its components.

2 Related work

In-context learning in VL. Frozen (Tsimpoukelli et al., 2021) is the first attempt for in-context learning in multimodality by leveraging a frozen GPT-like language model as the language backbone and mapping visual features to the language embedding space. Frozen sheds light on the feasibility of benefiting from the frozen LMs in VL modeling to learn a new task from a few examples in context. MAGMA (Eichenberg et al., 2021) is another encoder-decoder architecture for VL pre-training which showed that adding adaptor blocks between the frozen language model layers could further improve the performance for VL tasks in a few-shot scenario.

Other recent works (Yang et al., 2022; Alayrac et al., 2022; Zeng et al., 2022) follow the similar principle as the previous works to tackle in-context learning in VL modeling and achieve superior results by leveraging extremely large-scale models.

In this paper, we study a problem overlooked in prior work: we delve into the possibility of enabling in-context learning for VL tasks without relying on extensive scalability. Our focus lies in exploring the hypothesis: Is it feasible to transfer the in-context learning capability from the language domain to the VL domain?

Meta-learning in language modeling Large-scale language models have shown the capability to be trained on a new task if properly prompted with in-context examples, i.e., in-context learning. In this learning strategy, the language model is asked to generate the desired output, e.g., an answer in the question-answering task, which is prompted by a few data examples along with their corresponding supervision sampled from the training split, and the language model learns the task in context without performing any gradient updates. Although such training is highly data-efficient, its performance is far behind supervised fine-tuning. Therefore, inspired by (Vilalta and Drissi, 2002; Evgeniou and Pontil, 2004; Finn et al., 2017; Ruder, 2017), MetaICL (Min et al., 2022) proposes training the model for in-context learning as a kind of meta-learning. MetaICL meta-trained a gpt language model on a diverse set of natural language

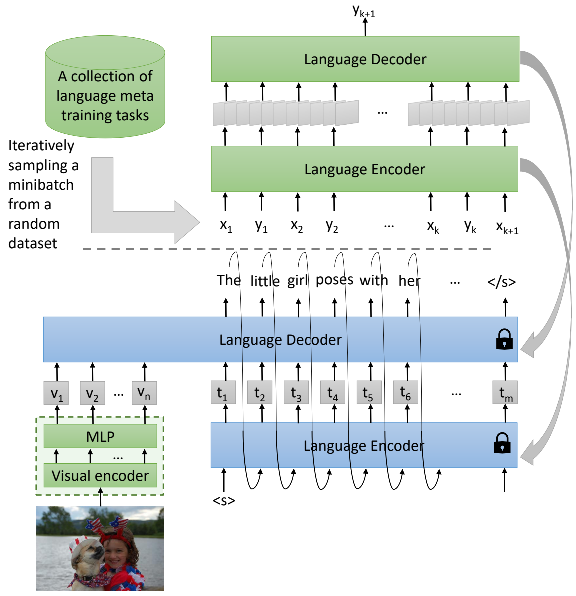

Figure 1: The training steps of MetaVL including meta-training the language encoder-decoder (above) and mapping the visual features into the language embedding space while keeping the meta-trained language encoder-decoder frozen (below).

tasks and datasets and showed that meta-training a language model in an in-context learning manner could significantly improve the in-context learning capability of the language model for a new task.

3 Approach

In this section, we first explain the existing meta-training procedure for language modeling and then introduce our proposed method for in-context learning in VL.

Meta-training in language modeling. MetaICL has shown that a language model that is meta-trained on a diverse set of tasks in an in-context learning setup is a strong few-shot learner. To meta-train an auto-regressive language model, in each iteration, a meta-learning task is randomly chosen from a collection of diverse meta-training language tasks, and $k+1$ data-label examples are randomly sampled from its training split. Then, the model is supervised by the concatenation of $(x_1, y_1, x_2, y_2, ..., x_{k+1})$ which will be fed as a single input to the model for predicting the label (y_{k+1}) as the training objective, i.e., the meta-training step aims to maximize:

$$P(y_{k+1}|x_1, y_1, \cdots, x_k, y_k, x_{k+1}) \quad (1)$$

During inference, the same in-context setup (k examples from the training) are sampled from a target dataset to be used as the $(x_1, y_1)(x_2, y_2) \cdots, (x_k, y_k)(x)$ and given to the model to predict the label y.

The meta-trained language model trained on a diverse set of natural language datasets has shown good performance for an unseen task when few data are given in context (Min et al., 2022).

MetaVL - a VL method with meta-learning knowledge for in-context learning. MetaVL has three main submodels including a meta-trained encoder-decoder and is being trained using Prefix Language Modeling (PrefixLM) (Wang et al., 2021). In the following, we discuss each submodel in detail.

Visual encoder and visual prefix. The visual encoder is defined as a function $V_e(x)$ that takes an image of x and outputs visual features. We extract the feature grid before the pooling layer $n \times D_v$ where n is the number of feature maps and D_v is the feature size of the visual encoder. Then, the output features can be viewed as a sequence of n visual tokens representing the image.

The visual encoder is followed by the visual prefix module that is defined as $V_p(x) \in D_v \times D_l$ which maps the visual features to language embedding space. This module is seeking to properly project the visual tokens into language tokens.

During the VL training, the parameters of both of these modules are trainable and are learned with different learning rates by back-propagation guided by the frozen language model.

Language encoder-decoder The meta-trained language encoder-decoder is used as the LM backbone and is frozen during the VL training process so the meta-trained language model preserves its few-shot capabilities. The language encoder encodes the text into text tokens represented by $t_1, t_2, ..., t_m$. Then, given the multimodal tokens (image and text) as $U = v_1, v_2, ..., v_n, t_1, t_2, ..., t_m$ the decoder is trained to reconstruct the corresponding text with a standard language modeling objective to maximize the following likelihood:

$$L(U) = \sum_{i=1}^{m} \log P(t_i | v_1, ..., v_n, t_1, ...t_{i-1}; \theta) \quad (2)$$

After the VL training, for learning a new VL task in-context, given a few examples from a new task with a new format, we concatenate k sampled data-label pairs from the training split along with one data from the val/test split to construct the prompt and feed it to the model for predicting the desired output. The entire process is visualized in Fig. 1.

4 Experiments

4.1 Datasets and Baseline

We use the dataset proposed in (Min et al., 2022) as the meta-training dataset for the language model and the COCO dataset (Lin et al., 2014) as the VL training dataset for MetaVL. The evaluation experiments are conducted on three datasets including VQA (Antol et al., 2015), OK-VQA (Marino et al., 2019), and GQA (Hudson and Manning, 2019). Frozen leveraged an internal GPT-like language model with 7 billion parameters as the backbone of their proposed model. As their model is not publicly available, we trained Frozen with GPT2-Medium as the frozen language model and consider it as our main baseline (Frozen_A) due to its model size. We also train a frozen with GPT-J 6B (The most similar GPT to Frozen) language model and obtained a close performance to the original Frozen model and use it as our second baseline denoted by Frozen_B.

4.2 Training and evaluation setting

Initially, We meta-train a GPT2-Medium LM on a collection of 142 meta-training language datasets with a learning rate of 1e-5 and a batch size of 8 using the setting named as "HR→LR with instructions (all)" where datasets with equal or greater than 10,000 training examples are used as meta-training tasks and the rest of the datasets are used as target tasks. The training is done on 8 NVIDIA RTX A6000 for 80,000 steps which took $\sim$ 6 hours. Then, we train MetaVL on the training split of COCO where we use a learning rate of 5e-5 and 2e-6 for the visual prefix and visual encoder, respectively, while the rest of the model parameters are frozen. We use a batch size of 32 and trained MetaVL using 4 NVIDIA RTX A6000 for 8 epochs which take $\sim$ 48 hours. Inference time depends on the numebr of shots varies from 2-5 hours for 0-3 shots on 5000 test examples. Our visual encoder is CLIP-RN50x16 (Radford et al., 2021) with a feature grid size of 144×3072 and our visual prefix is an MLP layer with a dimension of 3072×768. For in-context evaluation on VQA datasets, we randomly pick a specific number -n- of sampled data-label pairs, known as shots, from the training set and feed them to the model in-context followed by a single data from the val/test set. Fig. 2 provides some illustrative examples for the evaluation process.

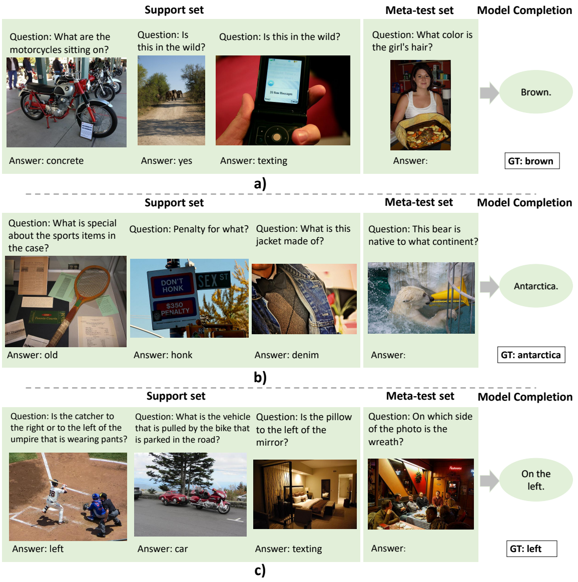

Figure 2: Qualitative examples of in-context learning from three datasets: a) VQA, b) OK-VQA, and c) GQA. For each example, there is also a task induction sentence of "please answer the question.".

To conduct the evaluation, we utilize a subset of 5,000 instances from the val/test dataset due to computational constraints. The generated output from the model is then compared against the expected answer, as established in previous studies. In cases where an exact match is not achieved, we employ a technique to identify the most closely related answer from a set of candidate answers (The set can be defined as a unique list of all answers in the training dataset). This involves computing the cosine similarity between the output's embedding and each candidate answer's embedding achieved by Sentence BERT (Reimers and Gurevych, 2019).

We then compare the selected output with the corresponding answer to determine the match. The training datasets for VQA, OK-VQA, and GQA contain approximately 3,000, 4,200, and 3,000 distinct answers, respectively. Furthermore, we performed an additional round of human evaluation on model's output without matching, and the findings are summarized in the appendix (Table 2). The human evaluation on a separate test set of 2000 examples aimed to delve deeper into instances where the model's output, while accurate, didn't precisely match the provided answer. Three such examples are presented in Fig 3, where the initial evaluation did not consider the prediction as correct, but it was deemed correct in the subsequent evaluation setting.

		Frozen_{A}	Frozen_{B}	MetaVL
	LM size	375M	7B	375M
Automatic evaluation	VQA	18.63	**34.07**	33.12
	OK-VQA	3.17	**11.97**	9.60
	GQA	13.86	25.76	**31.96**
Human evaluation	VQA	16.68	-	**35.09**
	OK-VQA	6.41	-	**19.22**
	GQA	19.96	-	**38.29**

Table 1: The performance of MetaVL compared with two baselines on 3-shot in-context learning. We report the performance of our re-implemented Frozen models.

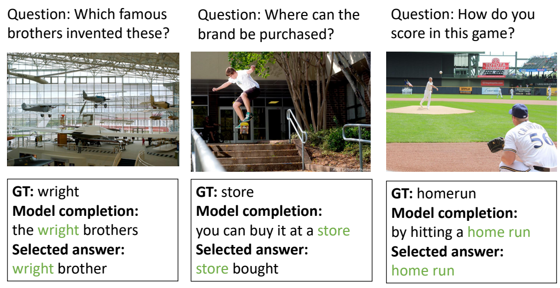

Figure 3: Three examples of VQA cases which The model's output, although correct, slightly differs from the ground-truth and selected answer from the candidate set.

4.3 Results and analysis

Quantitative analysis To evaluate MetaVL, we consider three common visual question-answering datasets including VQA, OK-VQA, and GQA. We compare MetaVL results with the mentioned two baselines in Table 1 for 3-shot in-context learning based on both automatic and human evaluation. According to the results, the performance of Frozen improves as its model size increases while MetaVL achieved competitive results in all three tasks. To further analyze how many image-text pairs are required to enable In-context learning for the VL task, we have trained MetaVl with 50 percent of training data and the results show that the performance slightly dropped but the model preserve its capability to learn from in-context data (Table 3).

The effect of the number of in-context shots According to Figure 4, in almost all settings, the performance of MetaVL is improving by increasing the number of shots which shows the model is gaining knowledge from the data in context. This result further gives us an illustration of the model's ca-

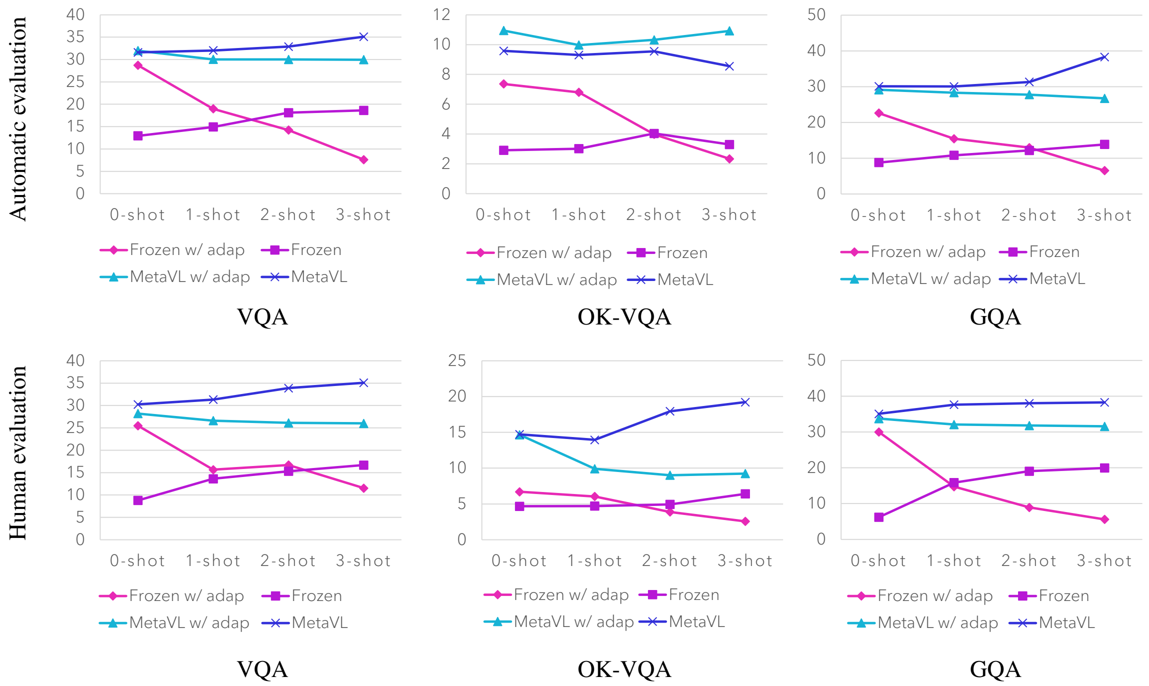

Figure 4: Automatic and human evaluation Accuracy of MetaVL and Frozen, w/ and w/o adaptors with 0-3 shots of in-context data.

pability to learn from the in-context examples supporting that MetaVL is benefiting from the meta-learning knowledge for in-context learning. The numbers on the graph are summarized in Table 2 in the appendix.

The effect of having adaptor layers in LM MAGMA claims that adding trainable adaptor layers and letting the LM slightly be trained during the VL training process is beneficial for in-context learning. Compared with Frozen, in addition to being trained on an x8 larger set of VL datasets, MAGMA also includes the training splits of the target datasets to its training set, while Frozen is adapted to an unseen new task in-context (in-context learning). We evaluated this method by adding adaptor layers to both Frozen and MetaVL and denoted the corresponding models by Frozen w/adap and MetaVL w/adap, respectively, in Fig. 4. Our results demonstrate that having a fully frozen language model in MetaVL could better preserve the in-context learning ability of the language model. It is also noticeable that adding adaptor layers improves the zero-shot performance of Frozen. We hypothesize that this improvement is due to getting a better vision and language alignment by letting both vision and language submodels be involved in the alignment process.

Qualitative analysis We provide some qualitative examples to better illustrate the performance of MetaVL for in-context learning in different VQA tasks. In Fig. 2, a few examples are provided which show the output of MetaVL for 3-shot in-context learning. More examples are presented in Appendix.

5 Conclusion

We investigate the feasibility of transferring meta-learning knowledge for in-context learning from resource-rich single modality to multimodality. We have shown that by leveraging a meta-trained language model in a VL model, we can transfer the ability of "learning to learn" in context to VL and it results in a strong VL few-shot leaner. With extensive experiments on three common VL datasets, we have shown that the in-context learning performance of MetaVL is superior compared with the baseline even when the size of our model is 20 times smaller.

6 acknowledgment

This work was supported by DARPA under agreement HR00112190130 and DARPA MSC program under agreement N660011924032. We would like to thank the reviewers for their feedback to improve this research work.

Limitations

While we have shown the potential of transferring in-context learning ability from a language model to VL tasks, the experiments in this paper are limited in two aspects. (1) We considered only the VQA task, which is limited in scope. It is unclear whether our method generalizes to other VL tasks. In fact, as most tasks in the VL domain take the form of visual question answering, it is less well-defined what would "cross-task generalization" entail in VL, compared to in NLP where (2) Due to computational limitations, we experiment with only a moderate-sized LM. It is unclear the performance of our method after scaling up.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Constantin Eichenberg, Sidney Black, Samuel Weinbach, Letitia Parcalabescu, and Anette Frank. 2021. Magma–multimodal augmentation of generative models through adapter-based finetuning. *arXiv preprint arXiv:2112.05253*.

Theodoros Evgeniou and Massimiliano Pontil. 2004. Regularized multi–task learning. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 109–117.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. 2023. Grounding language models to images for multimodal generation. *arXiv preprint arXiv:2301.13823*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

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.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2022. MetaICL: Learning to learn in context. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2791–2809, Seattle, United States. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212.

Ricardo Vilalta and Youssef Drissi. 2002. A perspective view and survey of meta-learning. *Artificial intelligence review*, 18(2):77–95.

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2021. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*.

Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. 2022. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3081–3089.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Andy Zeng, Adrian Wong, Stefan Welker, Krzysztof Choromanski, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, et al. 2022. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv preprint arXiv:2204.00598*.

A Appendix

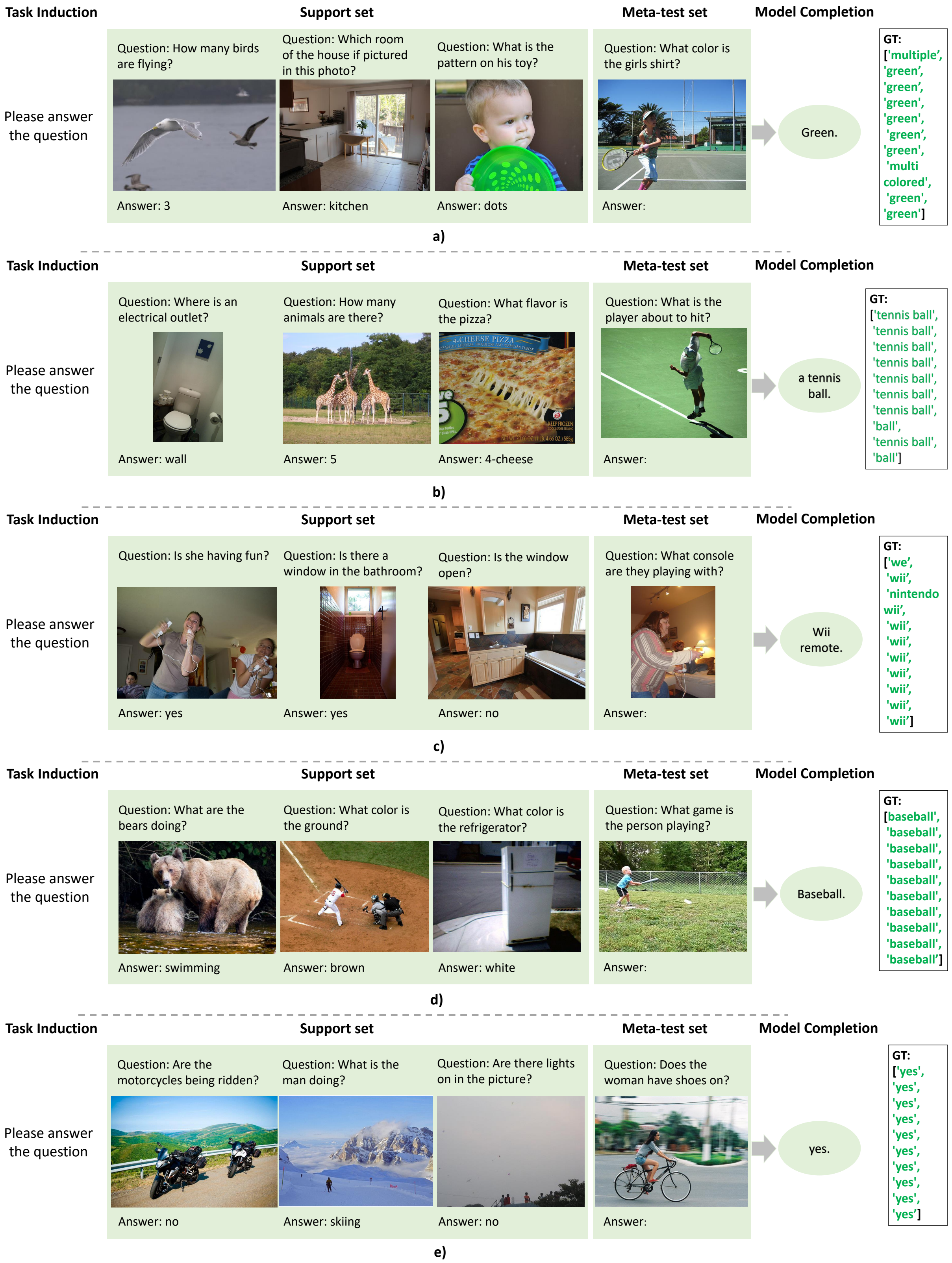

Figure 5: MetaVL success examples from VQA.

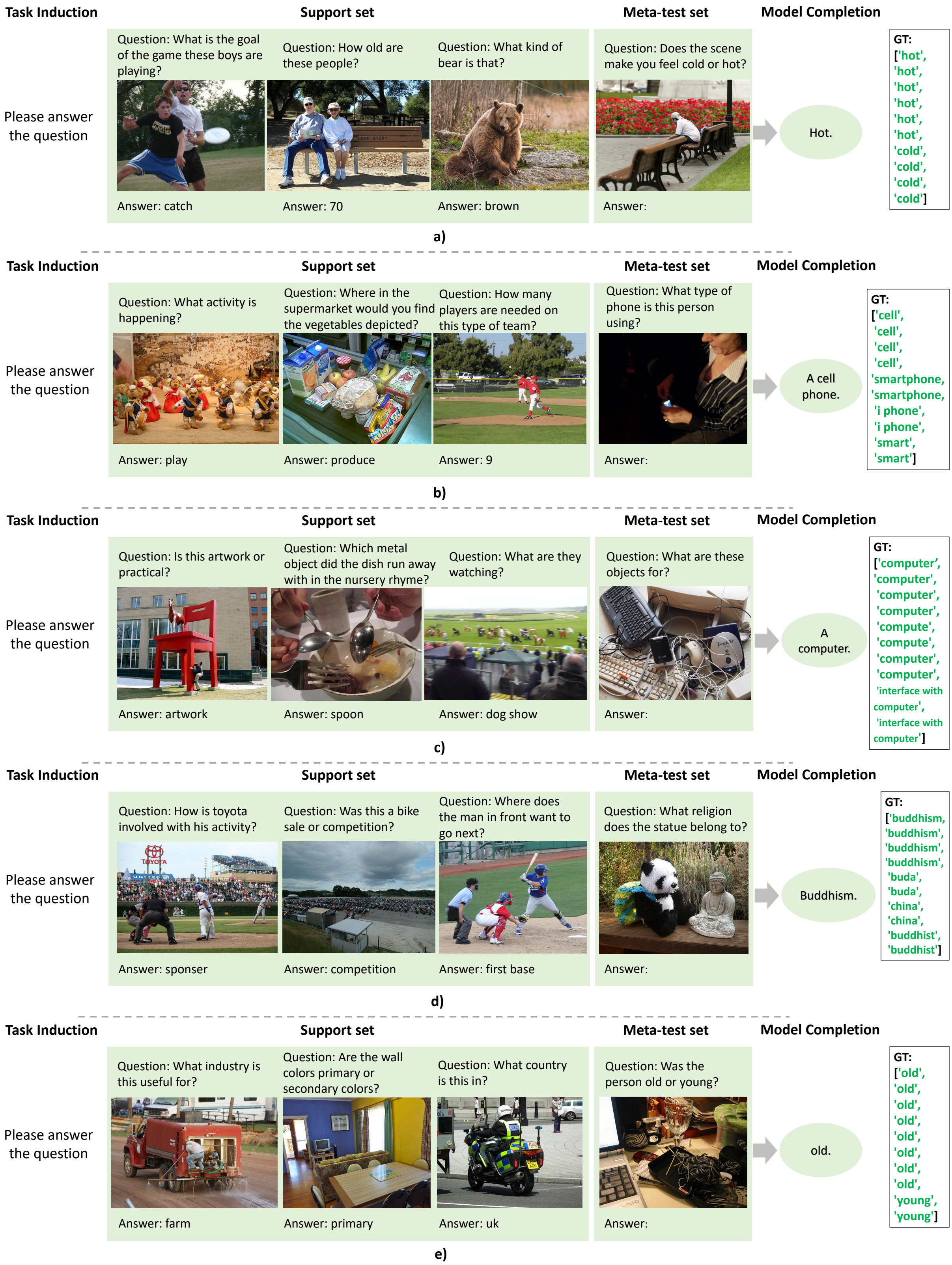

Figure 6: MetaVL success examples from OK-VQA.

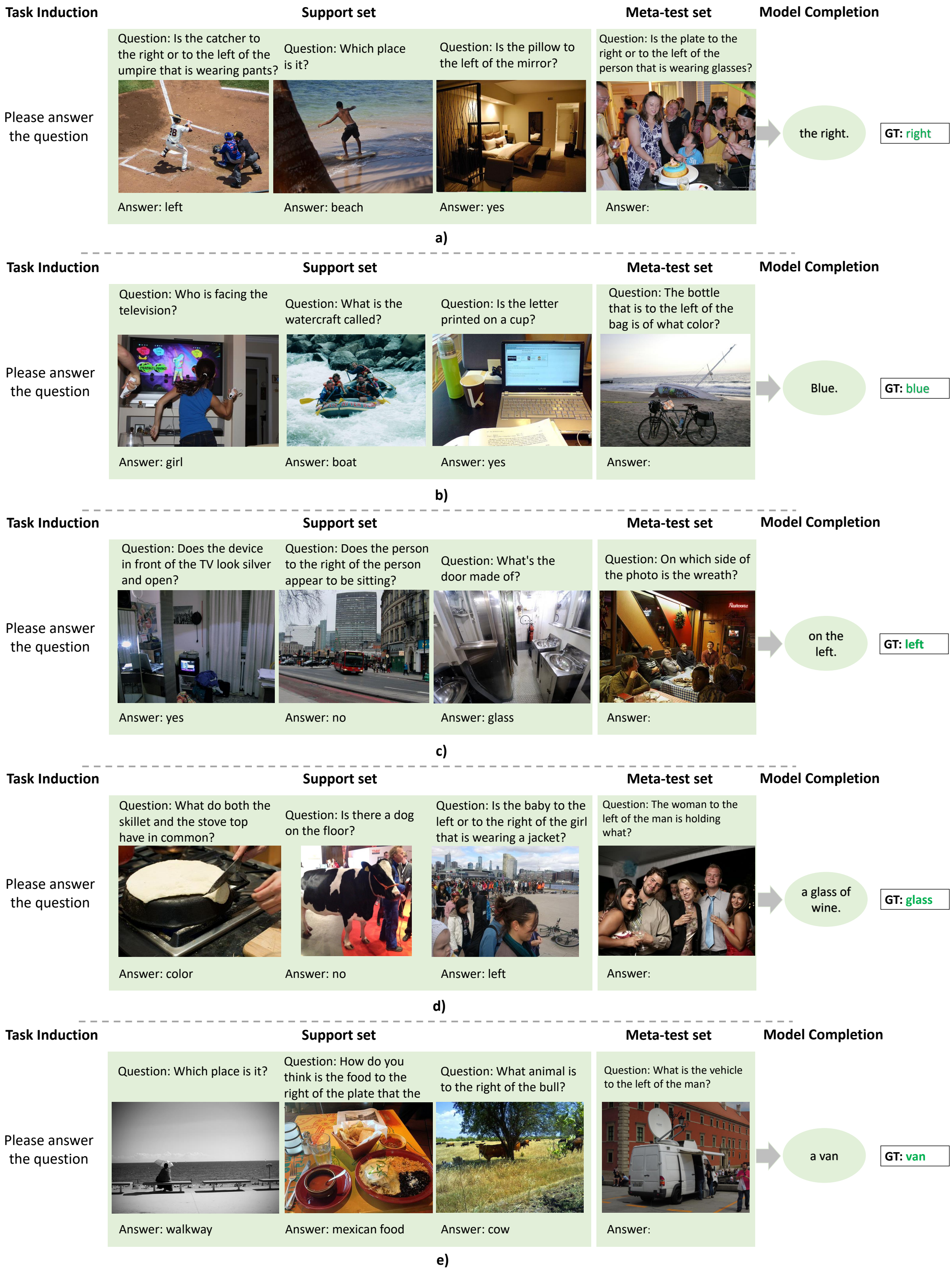

Figure 7: MetaVL success examples from GQA.

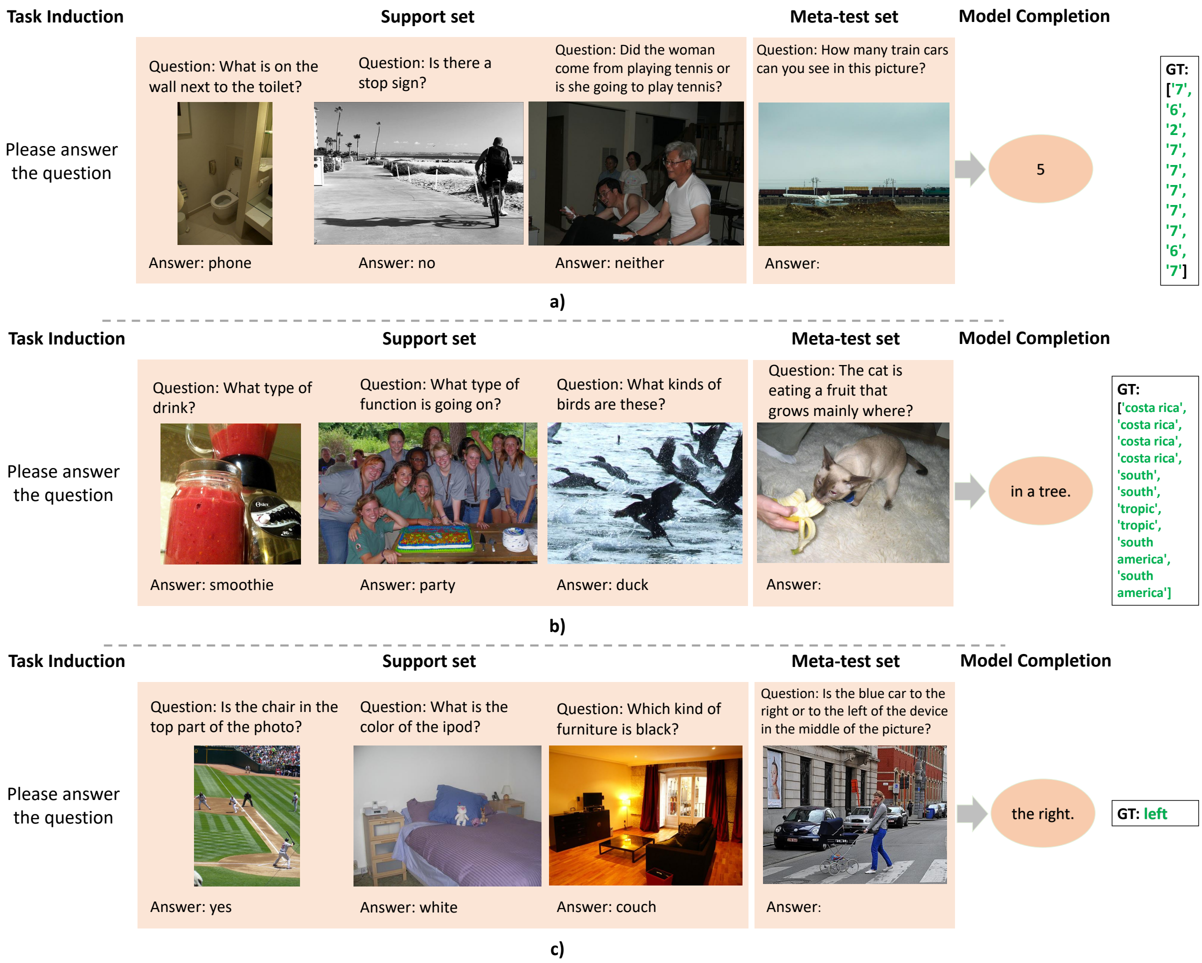

Figure 8: MetaVL failure examples from a) VQA, b) OK-VQA, and c) GQA.

model	Frozen_A w/ adap				Frozen_A				MetaVL w/ adap				MetaVL			
n-shot	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3
Automatic evaluation																
VQA	28.72	18.98	14.23	7.60	12.94	14.92	18.11	18.63	31.98	30.03	30.01	29.96	31.6	32.01	32.89	**33.12**
OK-VQA	7.36	6.30	3.98	2.34	2.91	3.02	4.04	3.30	**10.94**	9.97	10.32	10.92	9.58	9.30	9.55	9.60
GQA	22.62	15.44	12.96	6.54	8.80	10.81	12.17	13.86	29.12	28.31	27.78	26.74	30.10	30.05	31.32	**31.96**
Human evaluation																
VQA	25.49	15.66	16.70	11.53	8.79	13.62	15.31	16.68	28.20	26.61	26.12	26.01	30.24	31.33	33.89	**35.09**
OK-VQA	6.70	6.04	3.88	2.56	4.67	4.71	4.94	6.41	14.67	9.97	9.01	9.24	14.72	13.95	17.95	**19.22**
GQA	30.01	14.72	8.92	5.59	6.18	15.85	19.07	19.96	33.74	32.09	31.81	31.58	35.08	37.65	38.03	**38.29**

Table 2: Accuracy of MetaVL and Frozen, w/ and w/o adaptors with 0-3 shots of in-context data.

		MetaVL	$\text{MetaVL}_{50\%}$
Automatic evaluation	VQA	**33.12**	30.32
	OK-VQA	**9.60**	7.56
	GQA	**31.96**	27.77
Human evaluation	VQA	**35.09**	34.02
	OK-VQA	**19.22**	18.19
	GQA	**38.29**	35.66

Table 3: The performance of MetaVL was evaluated using the complete CoCo training dataset as well as a subset containing 50 percent of the CoCo training data. The experimental results indicate that even with the reduced training data, MetaVL maintains its capacity for in-context learning, albeit with a slight decrease in performance.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
6

☐ A2. Did you discuss any potential risks of your work?
Not applicable. Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☒ Did you use or create scientific artifacts?

Left blank.

☐ B1. Did you cite the creators of artifacts you used?
No response.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No response.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
No response.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
No response.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
No response.

☐ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
No response.

C ☑ Did you run computational experiments?

4.2

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
4.2

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
4.2

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
4.2

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
4.2

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?
Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.