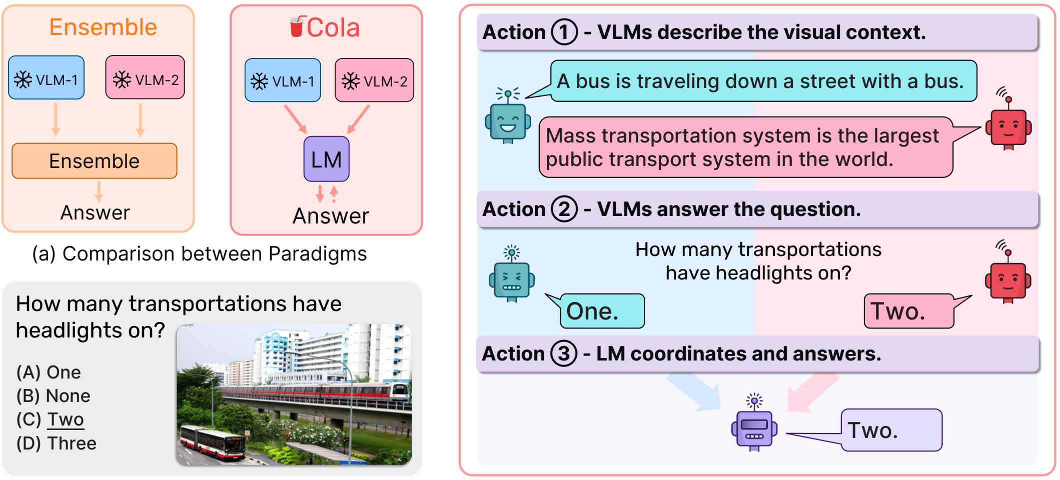
LANGUAGE MODELS ARE VISUAL REASONING COORDINATORS

Liangyu Chen[♥,*] **Bo Li**[♥,*] **Sheng Shen**[♣] **Jingkang Yang**[♥] **Chunyuan Li**[♠] **Kurt Keutzer**[♣]
Trevor Darrell[♣] **Ziwei Liu**[♥,✉]
[♥]S-Lab, Nanyang Technological University [♣]University of California, Berkeley [♠]Microsoft
{liangyu.chen, libo0013, ziwei.liu}@ntu.edu.sg

ABSTRACT

Visual reasoning demands multimodal perception and commonsense cognition of the world. Recently, multiple vision-language models (VLMs) have been proposed with excellent commonsense reasoning ability in various domains. However, how to harness the collective power of these complementary VLMs is rarely explored. Existing methods like ensemble still struggle to combine these models with the desired higher-order communications. In this work, we propose Cola [1], a novel paradigm that coordinates multiple VLMs for visual reasoning. Our key insight is that a language model (LM) can serve as an efficient coordinator to leverage the distinct and complementary capabilities of multiple VLMs. Extensive experiments demonstrate that our finetuning variant, Cola-FT, achieves state-of-the-art performance on outside knowledge VQA, visual entailment, and visual spatial reasoning tasks. Through systematic ablation studies and visualizations, we validate that a coordinator LM comprehends the instruction prompts and the separate functionalities of VLMs and then coordinates them to enable impressive visual reasoning capabilities.


Ensemble
VLM-1
VLM-2
Ensemble
Answer
Cola
VLM-1
VLM-2
LM
Answer
How many transportations have headlights on?
(A) One
(B) None
(C) Two
(D) Three
Action ① - VLMs describe the visual context.
A bus is traveling down a street with a bus.
Mass transportation system is the largest public transport system in the world.
Action ② - VLMs answer the question.
How many transportations have headlights on?
One.
Two.
Action ③ - LM coordinates and answers.
Two.

(a) Comparison between Paradigms

(b) Example Visual Reasoning Task: VQA

(c) Cola Solution to the Example VQA

Figure 1: We propose, Cola, using a coordinative language model for visual reasoning. Cola coordinates multiple pretrained VLMs based on the visual context and plausible answers they provide.

1 INTRODUCTION

Visual reasoning is a crucial task that demands models to not only comprehend and interpret visual information but also to apply high-level cognition to derive logical solutions (Johnson et al., 2017; Zakari et al., 2022; Małkiński & Mańdziuk, 2022b). Classic visual reasoners typically rely on

[*] Equal contribution. ✉Corresponding author.
[1] Code is available at https://github.com/cliangyu/Cola.

complex architectures (Yi et al., 2018; Mao et al., 2019; Yi et al., 2019) and are unable to generalize beyond the training dataset (Zellers et al., 2018; Park et al., 2020).

However, recent advancements in large pretrained models have shown that vision-language models (VLMs) exhibit impressive performance on visual reasoning tasks even under zero-shot settings (Wang et al., 2022; Li et al., 2022a). Meanwhile, language models (LMs) have also demonstrated robust zero-shot commonsense reasoning abilities on the natural language processing (NLP) applications (Brown et al., 2020; Chowdhery et al., 2022; Wei et al., 2022).

Consequently, several recent studies have attempted to combine such complementary VLMs and LMs for visual reasoning. For example, PICa (Yang et al., 2022) utilizes image captioning models to create textual prompts for GPT-3 (Brown et al., 2020), and adapts GPT-3 to solve the visual question answering (VQA) tasks in an in-context few-shot learning manner. Socratic Models (Zeng et al., 2022) allow VLMs and LMs to communicate via prompt engineering and emerging zero-shot multimodal reasoning capabilities. On the premise that current studies have focused on the interactions between heterogeneous models (*e.g.*, between VLM and LMs), in this work, we examine how to include homogeneous models (*e.g.*, multiple VLMs) with LMs in a coordinative paradigm. Inspired by the findings in CICERO (Meta et al., 2022) that LMs capture strong abilities in coordinating multiple agents, we propose Cola, a novel approach that utilizes an LM as the coordinator in between multiple VLMs.

Our key insight suggests that, *given multiple VLMs with different preferred patterns in describing the visual context and predicting plausible answers, an LM can coordinate and integrate their respective strengths efficiently and effectively*. We present two variants of Cola, namely Cola-FT and Cola-Zero, where FT corresponds to a finetuning approach and Zero stands for an in-context learning approach to adapt the coordinator LM for visual reasoning.

Systematic experiments demonstrate that Cola performs at the pinnacle of ability on outside knowledge VQA, visual entailment, and visual spatial reasoning tasks. Specifically, Cola-FT achieves state-of-the-art performance on A-OKVQA (Schwenk et al., 2022), e-SNLI-VE (Do et al., 2020), and VSR datasets (Liu et al., 2022a), even when compared with methods that adopt larger models or require more training computations. Besides, we conduct a thorough analysis to investigate how Cola recognizes each VLM's individual functionalities and then perform coordination behavior.

In summary, our contributions are as follows: **(1)** Cola, a novel paradigm that utilizes a language model as a coordinator between multiple VLMs to integrate their respective strengths for visual reasoning (§2). **(2)** Cola achieves state-of-the-art performance on a challenging suite of diverse visual reasoning datasets (§3.1). **(3)** Systematic analysis reveals how Cola comprehends the instruction prompts and the separate functionalities of VLMs and then coordinates them to capture impressive visual reasoning capabilities (§3.2, §3.3, §B.3).

2 Cola

2.1 🥤 Cola & Templates

An overview of Cola is shown in Figure 1c. We use OFA (Wang et al., 2022) and BLIP (Li et al., 2022a) as the VLMs and FLAN-T5 (Chung et al., 2022) as the LM. We first prompt each VLM to output captions and plausible answers independently. We then concatenate the instruction prompt, the question with choices, captions, and plausible answers to fuse all contexts for the LM to reason, coordinate, and answer.

General Prompt Template
Answer the following multiple-choice question by OFA and BLIP's description and their answers to the visual question. OFA and BLIP are two different vision-language models to provide clues. OFA's description: `<OFA caption>` BLIP's description: `<BLIP caption>` Q: `<Question>` OFA's answer: `<OFA answer>` BLIP's answer: `<BLIP answer>` Choices: `<Choices to the question>` A:

Table 1: LM prompt template. The LM is instructed to coordinate VLMs. Each question set defines *visual context*, *question with choices*, and *plausible answers*.

Image captioning gives important visual context to reason from. We first employ i^{th} VLM to describe each image respectively to get visual descriptions $c_i(v)$. We use ofa-large for OFA and blip-image-captioning-large for BLIP, both implemented by the Hugging Face Transformers library (Wolf et al., 2020).

Table 2: Overall performance on A-OKVQA, e-SNLI-VE and VSR datasets. The accuracy metric varies slightly in different datasets. In A-OKVQA, we report `val/test` accuracy, and `val` accuracy in e-SNLI-VE, `test` (zero-shot split) accuracy in VSR. We mark the best performance on each dataset with **bold font** and second-best with underlines.

Methods	Vision-language Model		Language Model			Accuracy ↑
	Model Spec.	FT ↓	Model Spec.	ICL ↓	FT ↓	
Outside Knowledge Visual Question Answering (A-OKVQA)						
VLC-BERT (Lu et al., 2019)	VL-BERT (118M)	20 ep.	-	-	-	- / 38.1
Unified-IO (Lu et al., 2022)	UNIFIED-IO (3B)	-	-	-	-	- / 45.2
PromptCap (Hu et al., 2022)	OFA (472M)	2 ep.	GPT-3 (175B)	-	-	- / 73.2
Img2Prompt (Guo et al., 2022)	BLIP (384M)	-	OPT (175B)	0-shot	-	42.9 / 40.7
Ensemble	BLIP+OFA (384M+472M)	-	-	-	-	56.6 / 54.9
🥤**Cola-Zero**	BLIP+OFA (384M+472M)	-	FLAN-T5 (11B)	0-shot	-	65.4 / 61.6
🥤**Cola-Zero**	BLIP+OFA (384M+472M)	-	FLAN-T5 (11B)	2-shot	-	70.4 / 66.5
🥤**Cola-FT**	BLIP+OFA (384M+472M)	-	FLAN-T5 (11B)	-	1 ep.	**77.7 / 74.0**
Visual Entailment (e-SNLI-VE)						
e-UG (Kayser et al., 2021)	UNITE (86M)	400 ep.	GPT-2 (117M)	-	400 ep.	79.5
OFA-X (Plüster et al., 2022)	OFA (472M)	10 ep.	-	-	-	80.9
Ensemble	BLIP+OFA (384M+472M)	-	-	-	-	48.8
🥤**Cola-Zero**	BLIP+OFA (384M+472M)	-	FLAN-T5 (11B)	0-shot	-	56.2
🥤**Cola-Zero**	BLIP+OFA (384M+472M)	-	FLAN-T5 (11B)	2-shot	-	57.8
🥤**Cola-FT**	BLIP+OFA (384M+472M)	-	FLAN-T5 (11B)	-	1 ep.	**81.6**
Visual Spatial Reasoning (VSR)						
VisualBERT (Li et al., 2019)	VisualBERT (110M)	100 ep.	-	-	-	54.0
LXMERT (Tan & Bansal, 2019)	LXMERT (110M)	100 ep.	-	-	-	63.2
ViLT (Kim et al., 2021)	ViLT (88M)	30 ep.	-	-	-	62.4
Ensemble	BLIP+OFA (384M+472M)	-	-	-	-	51.4
🥤**Cola-Zero**	BLIP+OFA (384M+472M)	-	FLAN-T5 (11B)	0-shot	-	55.8
🥤**Cola-Zero**	BLIP+OFA (384M+472M)	-	FLAN-T5 (11B)	2-shot	-	54.9
🥤**Cola-FT**	BLIP+OFA (384M+472M)	-	FLAN-T5 (11B)	-	1 ep.	**67.0**

Plausible answers by the VLMs to the question provide clues and patterns of VLMs for the LM to consider and coordinate. Similar to captioning, we prompt each i^{th} VLM using the image-question pair to get a plausible answer $\hat{a}_i(v, q)$. We use `ofa-large` for OFA and `blip-vqa-base` for BLIP. Following OFA, our prompt template varies by task category. For the VQA tasks, we leave the original question unchanged. For the visual entailment tasks, our prompt template is *" does the image describe* "`<text premise>`*" ?"*.

Prompt template is shown in Table 1. First, we design an instruction prompt for LM to understand the requirement to coordinate VLMs to answer the visual reasoning question. We then concatenate the captions from each VLM model, with the VLM identification labels in natural languages, such as *"OFA's description:* `<OFA caption>`*"*. Next, the question and its plausible answers provided by VLMs (with similar identification labels) are concatenated. We follow (Chung et al., 2022) to include the choices of question and *"A:"* to prompt for answers. More specific prompt templates on each dataset are provided in Appendix B.10.

2.2 🥤Cola-FT

Finetuning of Cola is initialized with FLAN-T5 (Chung et al., 2022) checkpoints. Given the question q based on the image v, the LM predicts the answer in the form of sequence $s_{v,q} = \mathrm{LM}(\mathrm{Prompt}(v, q))$.

Inference deploys the same prompt as Table 1 to align with finetuning. We resort to the greedy decoding strategy for conditional sequence generation at both finetuning and inference.

2.3 🥤Cola-Zero

In-context learning is an emerging ability of the LM models pretrained on documents of long-range coherence. By learning input and output format from demonstration, in-context learners learn to perform a downstream task simply by conditioning on a prompt consisting of input-output examples (Xie et al., 2021). FLAN-T5, finetuned on instruction prompts with examples, is capable of in-context few-shot learning and zero-shot learning (see Figures 4 and 5(b)).

Cola-Zero is the in-context few-shot/zero-shot learning variant of Cola, without finetuning. For in-context k-shot learning, we modify the prompt (Table 1) to include k input-output examples sampled from the training set. For zero-shot learning, the prompt remains the same as Table 1.

Question	What type of shot is the man hitting?	What appliance is next to an appliance that is highly decorated?	Does this image describe "puppy running after a stick in grass" ?	Does this image describe "The truck is away from the elephant" ?
OFA caption	tennis player hits a return to tennis player during their men's singles second round match at	a refrigerator covered in a variety of stickers.	a coyote is seen in this undated file photo. (credit: ktla	an elephant is loaded onto a truck in yangon. photo: afp
BLIP caption	a man in a blue shirt is playing tennis	a refrigerator with many pictures on it	a dog running through the grass in a field	a man riding a motorcycle with a truck behind him
Choices	['forehand', 'backhand', 'serve', 'dropshot']	['mixer', 'stove/oven', 'refrigerator', 'microwave']	['yes', 'maybe', 'no']	['yes', 'no']
OFA answer	backhand	stove/oven	yes	yes
BLIP answer	forehand	microwave	no	no
Cola-Zero answer	forehand	stove/oven	no	no
Cola-FT answer	forehand	stove/oven	maybe	no
Cola-FT answer (swapped VLM answer labels)	backhand	microwave	maybe	yes

Figure 2: Qualitative examples. The correct choices are underlined.

3 Experiments

The main quantitative results are then presented in Table 2. Next, we provide qualitative examples on how Cola integrates VLMs to provide final answer. Due to space limit, we leave the rest analysis and more details in Appendix B.

3.1 Overall Performance

In Table 2, we first observe that Cola-FT achieves state-of-the-art (SOTA) performance on all three datasets, with merely 1 epoch of finetuning and a medium-sized language model. In contrast, many previous SOTA methods require finetuning more epochs than Cola-FT (*e.g.*, VLC-BERT, PromptCap on A-OKVQA). Some also use much larger language models, such as GPT-3 (175B) (Brown et al., 2020) and OPT (175B) (Zhang et al., 2022). In addition, the lighter variant Cola-Zero also achieves comparable performance to most baseline methods through in-context few-shot and zero-shot learning, without training any model parameter.

3.2 Qualitative Examples

In Figure 2, we exhibit several qualitative examples. The leftmost example (a tennis player playing) demonstrates a case when captions are not informative to guide the LM for predictions. Between OFA and BLIP's plausible answers, the LM follows the answer of BLIP. In contrast, in the left example (an oven next to a fridge), again with trivial captions, the LM follows OFA's plausible answer instead.

The rightmost example presents the scenario of inconsistency between captions and answers. OFA describes the image as *"an elephant is loaded onto a truck in yangon."* Though, it agrees that *"the truck is away from the elephant"*. With Cola-FT, The LM coordinates OFA's correct caption and BLIP's correct answer to make a reasonable prediction.

Notably, we observe a scenario in that captions can be more informative than plausible answers to guide LM. The right example (a puppy running) presents an uninformative image. Though neither OFA nor BLIP succeeds to answer the question, the LM chooses to answer with "maybe" based on the given visual context. See Appendix B.11 and Appendix B.12 for more analysis on qualitative examples, including failure cases.

3.3 Coordination Analysis

Overall, Figure 3 validates the efficacy of Cola to coordinate VLMs. All the experiments use the same prompt template as in Table 1 unless otherwise stated. To validate the effectiveness of multi-VLM collaboration, we first ablate single-VLM variants of Cola-FT, shown as #1 and #2 from the top. As expected, both fall behind Cola-FT (#6) by a large margin. Next, we perturb caption labels by swapping the VLM caption labels at finetuning and evaluation (#3), specifically *”OFA’s description: ”* and *”BLIP’s description: ”*, by a chance of 50%. Under such settings, the LM fails to acquire the preferred patterns of VLM for captioning, though the overall visual context is preserved. The results underperform Cola-FT, which verifies that VLM caption labels improve Cola-FT performance. Notably, the VLM (plausible) answer labels are more important to the LM’s decision: a considerable gap exists between (#4) and Cola-FT. In #4, the LM fails to learn the separate functionalities of VLM. Naturally, we ask what if the LM can learn the patterns each VLM answers, but they cannot apply it at inference? We input correct VLM answer labels at finetuning and swap labels at evaluation (#5). Consequently, #5 falls behind Cola-FT with a smaller but still considerable margin. The results suggest that learning and applying the separate functionalities of VLMs is important for the coordinator LM to make predictions.

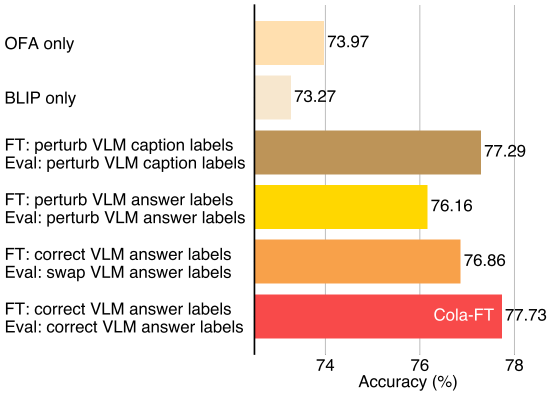


Figure 3: Ablation study results using (#1, #2 from top) single VLM, (#3, #4) perturbed VLM caption/answer labels at finetuning (FT), and (#5) swapped answer labels at evaluation (Eval). In #4, the coordination prior cannot be learned by the LM. In #5, the coordination prior can be learned by the LM, but cannot be properly applied at evaluation.

3.4 Scaling with More VLMs

By different top-k (k=5) decoding results from three identical (three OFA-base models, answers and captions may vary slightly, Cola achieved substantial performance gains over a single VLM or ensemble. The performance gap between ensemble baselines and Cola based on three different models (OFA-tiny, OFA-medium, and OFA-base) are even larger.

Methods	**A-OKVQA**	**e-SNLI-VE**
OFA-base (1)	45.76	52.60
OFA-base (2)	46.07	51.70
OFA-base (3)	45.73	52.33
Ensemble (majority voting)	44.79	52.71
Ensemble (average)	46.04	52.25
🥤 **Cola-Zero** (2-shot)	47.71	54.42
🥤 **Cola-FT**	48.85	56.92

Table 3: Performance of ensemble methods based on three identical models.

Methods	**A-OKVQA**	**e-SNLI-VE**
OFA-tiny	39.03	50.20
OFA-medium	42.45	51.04
OFA-base	45.76	52.60
Ensemble (majority voting)	46.71	53.94
Ensemble (average)	46.62	54.41
🥤 **Cola-Zero** (2-shot)	49.37	57.63
🥤 **Cola-FT**	54.26	63.68

Table 4: Performance of ensemble methods based on three different models.

4 Conclusions

In this paper we have proposed a novel paradigm for visual reasoning that harnesses the power of multiple VLMs. Experiments show that reasoning performance is substantially improved by LM finetuning or in-context learning. Our results provide a promising step towards building multi-component intelligent systems that capture multimodal reasoning capabilities in a human-like way.

References

Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Devi Parikh, and Dhruv Batra. Vqa: Visual question answering. *International Journal of Computer*

Vision, 123:4–31, 2015. 15, 17

J Alammar. Ecco: An open source library for the explainability of transformer language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pp. 249–257, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-demo.30. URL https://aclanthology.org/2021.acl-demo.30. 11

Saeed Amizadeh, Hamid Palangi, Alex Polozov, Yichen Huang, and Kazuhito Koishida. Neuro-symbolic visual reasoning: Disentangling. In *International Conference on Machine Learning*, pp. 279–290. PMLR, 2020. 17

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. 16

Alexander Bondarenko, Magdalena Wolska, Stefan Heindorf, Lukas Blübaum, Axel-Cyrille Ngonga Ngomo, Benno Stein, Pavel Braslavski, Matthias Hagen, and Martin Potthast. Causalqa: A benchmark for causal question answering. In *ACL*, 2022. 15, 16

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 2, 4, 16, 17

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021. 16

Zhuo Chen, Yufen Huang, Jiaoyan Chen, Yuxia Geng, Yin Fang, Jeff Z. Pan, Ningyu Zhang, and Wen Zhang. Lako: Knowledge-driven visual question answering via late knowledge-to-text injection. *ArXiv*, abs/2207.12888, 2022. 15, 17

Eunsol Choi, Daniel Hewlett, Jakob Uszkoreit, Illia Polosukhin, Alexandre Lacoste, and Jonathan Berant. Coarse-to-fine question answering for long documents. In *ACL*, 2017. 15, 16

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022. 2, 15, 17

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022. 2, 3, 12, 15, 17

Jonathan H Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. Tydi qa: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 2020. 16

Misha Denil, Alban Demiraj, and Nando De Freitas. Extraction of salient sentences from labelled documents. *arXiv preprint arXiv:1412.6815*, 2014. 11, 12

Thomas G Dietterich. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pp. 1–15. Springer, 2000. 16, 17

Virginie Do, Oana-Maria Camburu, Zeynep Akata, and Thomas Lukasiewicz. e-snli-ve: Corrected visual-textual entailment with natural language explanations. *arXiv preprint arXiv:2004.03744*, 2020. 2, 11

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. Eli5: Long form question answering. *arXiv preprint arXiv:1907.09190*, 2019. 16

Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*, 2020. 16

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 2021. 16

Dan Goldwasser and Dan Roth. Learning from natural instructions. *Machine learning*, 94(2):205–232, 2014. 16

Liangke Gui, Borui Wang, Qiuyuan Huang, Alex Hauptmann, Yonatan Bisk, and Jianfeng Gao. Kat: A knowledge augmented transformer for vision-and-language. *arXiv preprint arXiv:2112.08614*, 2021. 15, 17

Jiaxian Guo, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Boyang Li, Dacheng Tao, and Steven CH Hoi. From images to textual prompts: Zero-shot vqa with frozen large language models. *arXiv preprint arXiv:2212.10846*, 2022. 3

Yushi Hu, Hang Hua, Zhengyuan Yang, Weijia Shi, Noah A Smith, and Jiebo Luo. Promptcap: Prompt-guided task-aware image captioning. *arXiv preprint arXiv:2211.09699*, 2022. 3, 13

Drew A Hudson and Christopher D Manning. 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*, pp. 6700–6709, 2019. 15, 17

Anya Ji, Noriyuki Kojima, Noah Rush, Alane Suhr, Wai Keen Vong, Robert D Hawkins, and Yoav Artzi. Abstract visual reasoning with tangram shapes. In *EMNLP*, 2022. 15

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2901–2910, 2017. 1, 15, 17

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *ACL*, 2017. 16

Maxime Kayser, Oana-Maria Camburu, Leonard Salewski, Cornelius Emde, Virginie Do, Zeynep Akata, and Thomas Lukasiewicz. e-vil: A dataset and benchmark for natural language explanations in vision-language tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1244–1254, 2021. 3

Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pp. 5583–5594. PMLR, 2021. 3

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019. 16

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*, 2022a. 2, 12, 16, 17

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 3

Shuang Li, Yilun Du, Joshua B. Tenenbaum, Antonio Torralba, and Igor Mordatch. Composing ensembles of pre-trained models via iterative consensus. *ArXiv*, abs/2210.11522, 2022b. 16, 17

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014. 11

Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *arXiv preprint arXiv:2205.00363*, 2022a. 2, 11

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *arXiv preprint arXiv:2205.05638*, 2022b. 14

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 3

Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. *arXiv preprint arXiv:2206.08916*, 2022. 3

Mikołaj Małkiński and Jacek Mańdziuk. Deep learning methods for abstract visual reasoning: A survey on raven's progressive matrices. *arXiv preprint arXiv:2201.12382*, 2022a. 17

Mikołaj Małkiński and Jacek Mańdziuk. A review of emerging research directions in abstract visual reasoning. *arXiv preprint arXiv:2202.10284*, 2022b. 1, 17

Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum, and Jiajun Wu. The neuro-symbolic concept learner: Interpreting scenes words and sentences from natural supervision. *ArXiv*, abs/1904.12584, 2019. 2, 15, 17

Kenneth Marino, Xinlei Chen, Devi Parikh, Abhinav Gupta, and Marcus Rohrbach. Krisp: Integrating implicit and symbolic knowledge for open-domain knowledge-based vqa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14111–14121, 2021. 15, 17

John McCarthy et al. *Programs with common sense*. RLE and MIT computation center Cambridge, MA, USA, 1960. 16

Fundamental AI Research Diplomacy Team Meta, Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, et al. Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science*, 2022. 2, 16, 17

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. Metaicl: Learning to learn in context. *arXiv preprint arXiv:2110.15943*, 2021. 16

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. Cross-task generalization via natural language crowdsourcing instructions. *arXiv preprint arXiv:2104.08773*, 2021. 16

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022. 16

Jae Sung Park, Chandra Bhagavatula, Roozbeh Mottaghi, Ali Farhadi, and Yejin Choi. Visualcomet: Reasoning about the dynamic context of a still image. In *European Conference on Computer Vision*, 2020. 2, 17

Björn Plüster, Jakob Ambsdorf, Lukas Braach, Jae Hee Lee, and Stefan Wermter. Harnessing the power of multi-task pretraining for ground-truth level natural language explanations. *arXiv preprint arXiv:2212.04231*, 2022. 3, 13

Chen Qu, Liu Yang, Cen Chen, Minghui Qiu, W Bruce Croft, and Mohit Iyyer. Open-retrieval conversational question answering. In *ACM SIGIR*, 2020. 16

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021. 16, 17

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*, 2021. 17

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67, 2020. 15, 17

Nazneen Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. Explain yourself! leveraging language models for commonsense reasoning. In *ACL*, 2019. 16

Siva Reddy, Danqi Chen, and Christopher D Manning. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 2019. 16

Nils Reimers and Iryna Gurevych. 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, 11 2019. URL http://arxiv.org/abs/1908.10084. 11

Omer Sagi and Lior Rokach. Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1249, 2018. 16, 17

Shailaja Keyur Sampat, Maitreya Patel, Subhasish Das, Yezhou Yang, and Chitta Baral. Reasoning about actions over visual and linguistic modalities: A survey. *arXiv preprint arXiv:2207.07568*, 2022. 15, 17

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*, 2021. 16

Timo Schick and Hinrich Schütze. Exploiting cloze questions for few shot text classification and natural language inference. *arXiv preprint arXiv:2001.07676*, 2020. 16

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. *arXiv preprint arXiv:2206.01718*, 2022. 2, 11

Noam Shazeer and Mitchell Stern. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pp. 4596–4604. PMLR, 2018. 12

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020. 16

Md Arafat Sultan, Shubham Chandel, Ramón Fernandez Astudillo, and Vittorio Castelli. On the importance of diversity in question generation for qa. In *ACL*, 2020. 16

Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019. 3

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. Newsqa: A machine comprehension dataset. In *Rep4NLP@ACL*, 2016. 15, 16

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pp. 23318–23340. PMLR, 2022. 2, 12, 16, 17

Wenhui Wang, Hangbo Bao, Li Dong, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *ArXiv*, abs/2111.02358, 2021. 15, 17

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021. 12, 16

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022. 2

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://aclanthology.org/2020.emnlp-demos.6. 2

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment: A novel task for fine-grained image understanding. *ArXiv*, abs/1901.06706, 2019. 11, 15, 17

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. *arXiv preprint arXiv:2111.02080*, 2021. 3

Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *AAAI*, 2022. 2, 17

Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Joshua B. Tenenbaum. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. *ArXiv*, abs/1810.02338, 2018. 2, 15, 17

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. Clevrer: Collision events for video representation and reasoning. *ArXiv*, abs/1910.01442, 2019. 2, 17

Munazza Zaib, Wei Emma Zhang, Quan Z Sheng, Adnan Mahmood, and Yang Zhang. Conversational question answering: A survey. *Knowledge and Information Systems*, 2022. 15, 16

Rufai Yusuf Zakari, Jim Wilson Owusu, Hailin Wang, Ke Qin, Zaharaddeen Karami Lawal, and Yuezhou Dong. Vqa and visual reasoning: An overview of recent datasets, methods and challenges. *arXiv preprint arXiv:2212.13296*, 2022. 1, 15, 17

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6713–6724, 2018. 2, 15, 17

Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. In *Neural Information Processing Systems*, 2021. 15, 17

Andy Zeng, Adrian S. Wong, Stefan Welker, Krzysztof Choromanski, Federico Tombari, Aveek Purohit, Michael S. Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, and Peter R. Florence. Socratic models: Composing zero-shot multimodal reasoning with language. *ArXiv*, abs/2204.00598, 2022. 2, 16, 17

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022. 4

Fengbin Zhu, Wenqiang Lei, Chao Wang, Jianming Zheng, Soujanya Poria, and Tat-Seng Chua. Retrieving and reading: A comprehensive survey on open-domain question answering. *arXiv preprint arXiv:2101.00774*, 2021. 16

A Preliminaries

We formulate various visual reasoning tasks as a multi-class classification problem. Given an image $v \in \mathcal{V}$ and a question-like prompt $q \in \mathcal{Q}$, the reasoner is required to select an answer a from the candidate set $\mathcal{A} = \{a\}$. In the case that the reasoner outputs a text sequence $s_{v,q}$, we map s to a prediction $P(v, q) = \mathrm{sim}(T(s_{v,q}), T(\{a\}))$ where T transforms text sequences into text embeddings (we use a `all-mpnet-base-v2` model (Reimers & Gurevych, 2019) here), and sim denotes cosine similarity.

Ensemble Modeling is a prevalent method to combine multiple models' predictions in order to improve the overall performance (Figure 1a). One common practice is averaging over n models, given by:

$$P(v, q) = \frac{1}{n} \sum_{i=1}^{n} P_i(v, q), \tag{1}$$

where $P_i(v, q)$ denotes the prediction of the i^{th} model on input (v, q).

B Experiments Details

B.1 Datasets

Our experiments are conducted on a challenging suite of three diverse visual reasoning tasks, including outside knowledge VQA, visual entailment, and visual spatial reasoning. For each task, we select the following dataset respectively.

Augmented-OKVQA (Schwenk et al., 2022) (A-OKVQA) contains about 25k questions paired with both multiple choice (MC) answer options. Unlike most existing VQA datasets, the questions in A-OKVQA cannot often be answered by querying the knowledge base, but rather involve some type of commonsense reasoning and outside knowledge about the situation portrayed in the image.

e-SNLI-VE (Do et al., 2020) dataset is an extended version of SNLI-VE dataset (Xie et al., 2019), which contains about 190k question pairs and human-annotated natural language explanations for the ground-truth labels. The text premise provides a statement about the contents of the image. The task is to determine whether the statement is true or false based on the image content.

Visual Spatial Reasoning (Liu et al., 2022a) (VSR) consists of 65 spatial relations (*e.g.,* under, in front of, facing, *etc.*) of instances in images. VSR has more than 10k question pairs, associated with 6940 images from MS COCO (Lin et al., 2014).

B.2 Comparison Methods

State-of-the-art Methods are summarized into two broad categories, VLM alone, and VLM combined with LM. In Table 2, for a fair comparison, we detail the techniques (whether finetuning or in-context learning is required) used for training VLMs and LMs, and the number of training epochs.

Ensemble Modeling can be considered the most basic baseline for combining VLMs. It represents the base performance that the combination of VLMs can achieve on the target task when not processed additionally. We implement averaging ensemble (Equation (1)) of cosine similarity between VLM output and each choice of a question as our ensemble baseline.

B.3 Saliency Visualization

As shown in Figure 4, we visualize the importance of the input prompt tokens by input-gradient saliency feature attribution (Denil et al., 2014), implementing with `Ecco` (Alammar, 2021). The input tokens that are more relevant to predict the output token `"grass"` are highlighted in darker colors. In the given example, both Cola-FT and Cola-Zero predict the correct answer and find the relevant clues from visual context and plausible answers. Figure 4b shows that Cola-Zero attributes the output more to the instructions in the prompt template. This explains Cola-Zero's competitive performance, a consequence of FLAN instruction tuning (Wei et al., 2021). After finetuning, Cola-FT focuses more on the most informative parts of input: the question, choices, as well as VLMs' plausible answers.

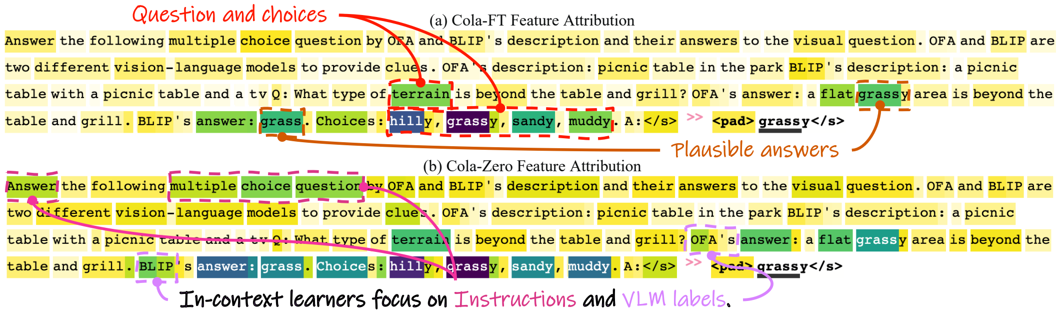

Question and choices
(a) Cola-FT Feature Attribution
Answer the following multiple choice question by OFA and BLIP ' s description and their answers to the visual question . OFA and BLIP are two different vision - language models to provide clues . OFA ' s description : picnic table in the park BLIP ' s description : a picnic table with a picnic table and a tv Q: What type of terrain is beyond the table and grill? OFA ' s answer : a flat grassy area is beyond the table and grill . BLIP ' s answer : grass . Choices : hilly , grassy , sandy , muddy . A: </s> >> <pad> grassy</s>
Plausible answers
(b) Cola-Zero Feature Attribution
Answer the following multiple choice question by OFA and BLIP ' s description and their answers to the visual question . OFA and BLIP are two different vision - language models to provide clues . OFA ' s description : picnic table in the park BLIP ' s description : a picnic table with a picnic table and a tv Q: What type of terrain is beyond the table and grill? OFA ' s answer : a flat grassy area is beyond the table and grill . BLIP ' s answer : grass . Choices : hilly , grassy , sandy , muddy . A: </s> >> <pad> grassy</s>
In-context learners focus on Instructions and VLM labels.

Figure 4: Visualization of input token saliency. We visualize the relevancy between input tokens and the output token "grass" by feature attribution (Denil et al.). The more salient tokens are highlighted in darker boxes. Cola-FT focuses on the question, choices, and VLMs' plausible answers in (a). While as shown in (b), Cola-Zero pays extra attention to instructions and VLM labels, as a consequence of instruction finetuning (Chung et al.).

B.4 Scaling Model Size

We conduct experiments on scaling model size to see if there are ramifications when operating at a larger scale. Figure 5(a) reveals that Cola-FT performance increases as the LM (FLAN-T5) model size increases. Notably, Cola-FT/small, with only 80M parameters, could achieve about 65% MC accuracy on A-OKVQA validation set, which is far beyond our baseline methods (around 55%). Cola-Zero, under the in-context learning paradigm, achieves competitive performance when the model grows to a billion-parameter scale. This observation on Cola-Zero can be regarded as a proof-of-concept that potentially reveals Cola-Zero's emerging abilities (inherited from FLAN-T5 (Chung et al., 2022)) on visual reasoning tasks at a relatively large scale.

B.5 Low-data Finetuning

We conduct experiments on different data scales to verify Cola's performance varying from zero-shot to full-shot under in-context learning and full-finetune paradigm. As shown in Figure 5(b), with Cola-Zero, few-shot exemplars substantially improve performance compared to zero-shot learning. As Chung et al. (2022); Wei et al. (2021) pointed out, exemplars potentially help the model better understand the output format and understand the instructions in Table 1.

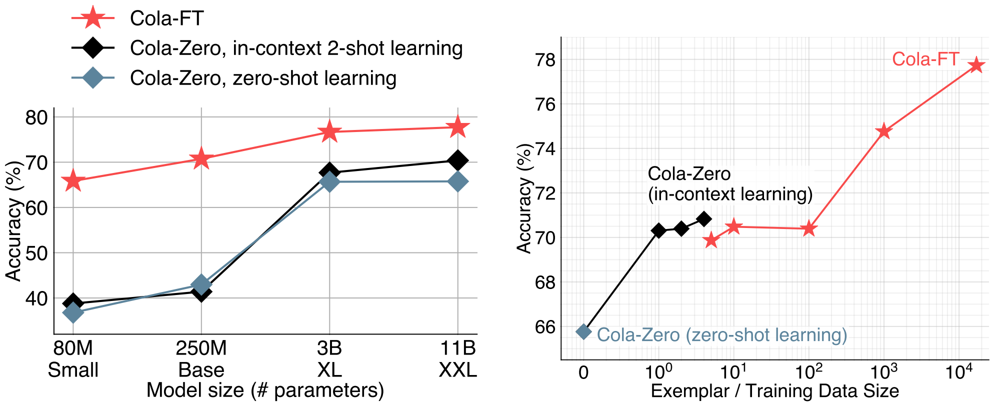
We also observe Cola-FT's substantial performance gain when finetuning shots increase to 1000. Cola-FT keeps improving till the inclusion of the whole training set.

B.6 Finetuning Details

We adopt pretrained BLIP (Li et al., 2022a)[2] and OFA (Wang et al., 2022)[3] as VLMs, and freeze their parameters without updating. The finetuning only happens on the language model part. The training set of each dataset is used for finetuning. We use the whole training set unless otherwise specified in low-data finetuning discussion.

We use an AdaFactor optimizer (Shazeer & Stern, 2018) at the learning rate of 1e-4 for all Cola-FT experiments. The batch size is by default set to 16, though we find Cola-FT insensitive to batch size. We finetune and evaluate the models on NVIDIA V100 or A100 GPUs. The finetuning ranges from 1 hour to about 15 hours, varying by the dataset.

[2] BLIP: https://github.com/salesforce/BLIP
[3] OFA: https://huggingface.co/OFA-Sys



(a) Model size ablation.

(b) Low-data settings.

Figure 5: (a) Cola performances versus the LM (FLAN-T5) sizes, on A-OKVQA validation set. Cola-FT is applicable with small models, but Cola-Zero is an emerging ability on larger models only. (b) Low-data Cola-FT and Cola-Zero performances on A-OKVQA validation set. Cola-Zero for in-context few-shot learning outperforms zero-shot learning by a large margin, being on par with low-data Cola-FT without finetuning. The X-axis is manipulated to exhibit zero-shot learning.

Following the common experiment protocols, we employ a teacher forcing and greedy decoding strategy for finetuning.

B.7 Evaluation Details

As specified, we use the validation or test set multiple choice accuracy as the evaluation metric. In A-OKVQA, we report `val`/`test` accuracy, and `val` accuracy in e-SNLI-VE, `test` (zero-shot split) accuracy in VSR. For simplicity and consistency, we evaluate ablation experiments on A-OKVQA validation set. Following the common experiment protocols (Hu et al., 2022; Plüster et al., 2022), we report the single run results for performance comparison.

The exemplars at the inference of Cola-Zero are randomly sampled from the training set, i.e., supposedly help the LM learn the input data distribution and output format but do not leak relevant information to the evaluation question.

B.8 A-OKVQA Direct Answer Results

In addition to MC accuracy, we present the direct answer (DA) accuracy of models on the A-OKVQA validation set in Tables 5 and 6.

	FLAN-T5-Small	FLAN-T5-Base	FLAN-T5-XL	FLAN-T5-XXL
Cola-FT	56.5	60.6	64.1	65.4
Cola-Zero (2-shot)	30.3	34.6	57.6	61.0
Cola-Zero (0-shot)	28.6	36.0	55.0	59.3

Table 5: A-OKVQA validation set DA performance. Extension of Figure 5(a).

	1-shot	2-shot	3-shot	4-shot
Cola-Zero	60.2	61.0	60.7	59.2

Table 6: Cola-Zeroin-context few-shot learning DA performance on A-OKVQA validation set. Extension of Figure 5(b).

B.9 Parameter-efficient Finetuning

To further reduce the computation cost in model adaptation, we explored parameter-efficient finetuning (PEFT) techniques to reduce finetuning parameter counts. Specifically, we use $(\texttt{IA})^3$ (Liu et al., 2022b), which finetunes an overhead of 1 million parameters, equivalent to 0.01% of the full parameters of FLAN-T5-XXL.

	Accuracy	**# Finetuning Params**
Finetuning	77.73	11B (100%)
PEFT, $(\texttt{IA})^3$	63.76	1M (0.01%)

Table 7: $(\texttt{IA})^3$ (Liu et al., 2022b) parameter-efficient tuning (PEFT) performance. We finetune a FLAN-T5-XXL model on the A-OKVQA training set and evaluate it on the A-OKVQA validation set.

Compared to full finetuning, $(\texttt{IA})^3$ requires more iterations to converge. The performance of a $(\texttt{IA})^3$ finetuned FLAN-T5-XXL model is on par with a fully finetuned FLAN-T5-Small (80 million parameters) counterpart (Figure 5(a)). Notably, the former is associated with more computation and memory footprint as a consequence of more parameters in the forward pass.

B.10 Extended Prompt Templates

Across three datasets, the prompt template is roughly the same, with minor differences mainly in the format of the questions and choices. We list the prompt templates adopted in A-OKVQA and e-SNLI-VE/VSR in Table 8 and Table 9, respectively.

A-OKVQA Prompt Template
Answer the following multiple-choice question by OFA and BLIP's description and their answers to the visual question. OFA and BLIP are two different vision-language models to provide clues. OFA's description: `<OFA caption>` BLIP's description: `<BLIP caption>` Q: `<Question>` OFA's answer: `<OFA answer>` BLIP's answer: `<BLIP answer>` Choices: `<Choices to the question>` A:

Table 8: A-OKVQA prompt template for the LM. The LM is instructed to coordinate VLMs. Each question set defines *visual context*, *question with choices*, and *plausible answers*.

B.11 Extended Qualitative Examples

In this section, we provide more qualitative examples on A-OKVQA (Figure 6), e-SNLI-VE (Figure 7), and VSR (Figure 8) datasets.

Due to the large span of the three figures, for better visibility, we put the detailed description directly in each figure's caption part. We illustrate how Cola-FT and Cola-Zero process the VLMs answers in each example.

Overall, in these examples, we can observe that even if BLIP and OFA provide wrong answers, Cola can still present the correct answer based on the captions provided by OFA and BLIP, as well as the choice set. This may illustrate how Cola amazingly accompanises visual reasoning tasks via coordinating BLIP and OFA.

e-SNLI-VE / VSR Prompt Template

Answer the following multiple-choice question by OFA and BLIP's description and their answers to the visual question. OFA and BLIP are two different vision-language models to provide clues.

OFA's description: `<OFA caption>`
BLIP's description: `<BLIP caption>`

Q: does the image describe `<hypothesis>` ?

OFA's answer: `<OFA answer>`
BLIP's answer: `<BLIP answer>`

e-SNLI-VE Choices: [yes, no, maybe]
VSR Choices: [yes, no]

A:

Table 9: e-SNLI-VE/VSR prompt template for the LM. The LM is instructed to coordinate VLMs. Each question set defines *visual context*, *hypothesis*, and *plausible answers*.

B.12 Failure Cases

In Figure 9, we provide a few failed cases to analyze the specific behavior of Cola.

The leftmost example's correct answer is *kayaking*, but there are no hints from OFA and BLIP's answers and captions. Therefore Cola-Zero incorrectly provides the answer *OFA* without sufficient information as hints, while surprisingly Cola-FT answered correctly from OFA's *boating* answer.

The left example again has insufficient information from captions. While BLIP answers *no* and OFA answers *yes*, Cola-FT chooses to answer *maybe*, which looks natural but unfortunately picks the wrong choice.

The right example's captions contain enough information this time. But both Cola-FT and Cola-Zero are misled by BLIP's wrong answer *no parking*.

The rightmost example also has insufficient information from captions. In this situation, Cola has no choice but to believe either BLIP or OFA's answer, but it mistakenly prefers BLIP's wrong answer.

C Related Works

C.1 Visual Reasoning

Beyond unimodal reasoning tasks such as question answering (QA) (Trischler et al., 2016; Choi et al., 2017; Zaib et al., 2022; Bondarenko et al., 2022), visual reasoning extends high-level cognition to visual domains, requiring an intelligent agent to derive rational solutions (Johnson et al., 2017; Hudson & Manning, 2019; Sampat et al., 2022; Zakari et al., 2022; Ji et al., 2022). Several tasks have been introduced to address visual reasoning, such as VQA (Agrawal et al., 2015), in which models are expected to provide answers to questions related to an image, and visual entailment (Xie et al., 2019), where the model is required to determine if a text description is consistent with the visual content provided.

Classic visual reasoning methods employ an image encoder and a text encoder, along with a reasoning block that utilizes attention mechanisms (Zellers et al., 2018; 2021; Wang et al., 2021), neuro-symbolic methods (Yi et al., 2018; Mao et al., 2019), or external knowledge (Marino et al., 2021; Gui et al., 2021; Chen et al., 2022).

Recent progress in large pretrained models has led to the development of LMs that capture exceptional commonsense reasoning capabilities (Raffel et al., 2020; Chung et al., 2022; Chowdhery et al., 2022). These LMs can potentially replace the reasoning module in visual reasoning tasks, and

LMs' lack of perception can be compensated by incorporating multiple VLMs trained on different domains (Radford et al., 2021; Wang et al., 2022; Li et al., 2022a). However, there is still a lack of research on how to harness the collective power of these separate VLMs for visual reasoning tasks.

C.2 Model Ensemble

Model ensemble is a powerful machine learning technique that combines the predictions of multiple models to improve the overall performance of a given task (Dietterich, 2000). The variance and bias of the final predictions decrease, resulting in a more robust and accurate model (Sagi & Rokach, 2018). To this end, common methods include averaging, voting, weighting the predictions based on model performance, or stacking the models.

Ensemble methods have been challenging for visual reasoning, where a simple combination is not applicable to heterogeneous models as their inputs and outputs vary. To address the issue, Socratic Models (SMs) (Zeng et al., 2022) use prompt engineering to guide the heterogeneous pretrained multimodal models through natural language discussions. With a similar goal, Li et al. propose a closed-loop iterative consensus optimization method to utilize the strengths of individual models. However, previous methods do not fully adapt to the intrinsic patterns of different models, particularly in the visual reasoning scenario. Recent studies, such as CICERO (Meta et al., 2022), have shown that LMs possess strong social intelligence in coordinating multiple agents, which inspires us to reorganize pretrained mixed-modal models with a focus on adapting LMs.

C.3 Finetuning Large Language Models

Large language models (Brown et al., 2020; Ouyang et al., 2022; Bommasani et al., 2021) pretrained on massive amounts of unstructured data have gradually demonstrated great performance by finetuning on additional task-specific instances. Finetuning a large language model can be considerably more sample efficient than re-training from scratch, although acceptable performance may still require a considerable quantity of data (Stiennon et al., 2020). Recent works have finetuned task-specific models that demonstrate amazing capabilities in many real-world applications, such as Copilot for program synthesis (Chen et al., 2021).

C.4 Instructions-based Learning

Recent advances in the capabilities of language models have piqued researchers' curiosity in the field of instruction-based learning (Goldwasser & Roth, 2014; McCarthy et al., 1960; Schick & Schütze, 2020; Gao et al., 2020). The core of instruction-based learning is to explore the knowledge of the language model itself. In the contrast to prompt learning to stimulate the language model's ability to complete blanks, instruction tuning more focuses on activating the language model's comprehension by giving obvious instructions to models and expecting correct feedback. Earlier work (Mishra et al., 2021) finetune BART (Lewis et al., 2019) using instructions and few-shot examplars in question answering, text classification, and text modification. Their findings suggest that few-shot instruction tuning improves performance on unseen tasks. (Min et al., 2021) finetunes GPT-2 Large and also observes that few-shot examplar instruction tuning could improve performance. (Sanh et al., 2021) finetunes T5-11B with more diverse instruction templates and observe similar improvements in zero-shot learning. More recent work (Wei et al., 2021) performs large-scale experiments with a 137B FLAN-T5 model and instruction-tune it on over 60 datasets verbalized via instruction templates. They observe FLAN-T5 substantially improves over zero-shot GPT-3 (175B) on 20 of 25 evaluation datasets. OpenAI also releases InstructGPT (Ouyang et al., 2022) based on GPT-3 (Brown et al., 2020), it makes use of human annotations to steer desired model behavior through both instruction tuning and reinforcement learning of human feedback. They discover that InstructGPT is favored by humans over unmodified GPT-3.

C.5 Visual Reasoning

Beyond the uni-modal reasoning tasks such as question answering (QA) (Trischler et al., 2016; Joshi et al., 2017; Choi et al., 2017; Reddy et al., 2019; Rajani et al., 2019; Fan et al., 2019; Qu et al., 2020; Clark et al., 2020; Sultan et al., 2020; Geva et al., 2021; Zhu et al., 2021; Zaib et al., 2022; Bondarenko et al., 2022), visual reasoning requires models to not only understand and interpret visual information but also to apply high-level cognition to derive rational solutions (Johnson et al., 2017; Hudson & Manning, 2019; Amizadeh et al., 2020; Małkiński & Mańdziuk, 2022a;b; Sampat et al., 2022; Zakari et al., 2022). Several tasks have been introduced to address visual reasoning, such as visual question answering (VQA) (Agrawal et al., 2015), in which models are expected to provide answers to questions related to an image, and visual entailment (VE) (Xie et al., 2019), where the model is required to determine the similarity or relationship between a given image and a description. Classic visual reasoning methods have employed an image encoder and a text encoder, along with a reasoning block that utilizes attention mechanisms (Zellers et al., 2018; Park et al., 2020; Zellers et al., 2021; Wang et al., 2021), neuro-symbolic methods (Yi et al., 2018; Mao et al., 2019; Yi et al., 2019), or external knowledge (Marino et al., 2021; Gui et al., 2021; Chen et al., 2022) to perform reasoning.

Recent progress in large pre-trained models has led to the development of language models (LMs) that possess exceptional commonsense reasoning capabilities (Raffel et al., 2020; Chung et al., 2022; Chowdhery et al., 2022; Rae et al., 2021). These models can potentially replace the reasoning block in visual reasoning tasks, and LMs' lack of perception can be compensated by incorporating multiple vision-language models (VLMs) trained on different domains (Radford et al., 2021; Wang et al., 2022; Li et al., 2022a). For example, PICa (Yang et al., 2022) converts the image into captions that GPT-3 (Brown et al., 2020) can understand, and adapts GPT-3 to solve the VQA task in a few-shot manner by providing a few in-context VQA examples. However, there is still a lack of research on how to harness the collective power of these complementary VLMs for visual reasoning tasks.

C.6 Model Ensembling

Model ensembling is a powerful machine learning technique that combines the predictions of multiple models to improve the overall performance of a given task (Dietterich, 2000). Classic model ensembling methods include simple averaging, weighting the predictions based on model performance, and stacking the models. By combining the predictions of multiple models, ensembling can reduce the variance and bias of the final predictions, resulting in a more robust and accurate model (Sagi & Rokach, 2018). Ensemble methods have been shown to perform well in a wide range of tasks, including image classification, natural language processing, and time series forecasting. However, when it turns to multimodal tasks such as visual reasoning, a simple combination is not applicable to heterogeneous models as their inputs and outputs vary.

To address the problem, Socratic Models (SMs) (Zeng et al., 2022) use prompt engineering to guide the heterogeneous pre-trained multimodal models through multimodal discussions to combine their diverse knowledge. With a similar goal, (Li et al., 2022b) proposes a closed-loop iterative consensus optimization method to utilize the strengths of individual models. However, previous methods do not fully explore the potential of a centralized solution or adapt to the separate functionalities of different models, particularly in the visual reasoning scenario. Recent studies, such as CICERO (Meta et al., 2022), have shown that language models possess strong capabilities in coordinating multiple agents, which inspires us to reorganize pre-trained multimodal models with a focus on the language models.

D Looking Back and Forward: a Concluding Dialogue

David and Brian (pseudo names) are two graduate students studying visual reasoning and have interests in recent emergent large language models. Over a cup of coffee, they sit down together to discuss the advantages and caveats that language models could bring to visual reasoning.

David: I've noticed that recent emergent large language models, aka LLM, have a wide range of abilities, including the capability to generate program code, analyze sentiments, and even solve math problems via logical reasoning.

Brian: That's incredible, and I've seen relevant papers on it. The LLM possesses these abilities fully from unsupervised pretraining linguistic data and could be stimulated by finetuning on specific downstream tasks. Researchers call these emerging abilities, and I've seen discussions related to multi-step reasoning, chain-of-thought, and so on.

David: Interesting! Similar models also exist in the computer vision field, such as CLIP, DALL-E, etc. They are also very strong at capturing visual signals and can achieve very good results on many vision tasks. Those large-scale models have demonstrated overwhelming capability, and people seem to have entered the era of foundation models.

Brian: It's really amazing, therefore I'm questioning whether these models of various modalities may be combined to accomplish something. At the moment it doesn't look like the language model has a way to process visual information well, and the visual language model doesn't seem to be a strong reasoner of language information. Will they have more incredible powers if they could compliment one another? One might also avoid having to repeatedly train new foundation models.

David: What you said is interesting; I saw a similar work with this idea called Socratic Models, in which various modalities (e.g. visual language models, and language models) models were allowed to *communicate* via language prompts, it's like the models ask questions to each other and finally let one output final answer.

Brian: It's amazing how fast everyone's research has progressed. So have they tried using VLM to capture visual information and provide caption prompt, and LLM for further inference like a reasoner? This is very suitable for VQA-related tasks that require powerful reasoning ability.

David: Not yet, I see they provide some examples about this paradigm, but not including VQA task. Maybe we can think deeper in this direction, I think it makes a lot of sense to combine multiple models, naturally different models may have their own special preferred patterns. Naturally, VLM is more suitable for processing visual information, and LLM is suitable for processing language information. And even for homogeneous models, like multiple VLMs, they may also have preferred patterns, which may be determined by the dataset they are trained on and specific training strategy.

Brian: Indeed, combining multiple pretrained models would make them more powerful. Ensemble modeling is a good way, however, it appears rather basic and unexplainable. The combination of multiple VLMs with LLMs is intriguing, in a Socratic Models way, and perhaps LLMs will learn to integrate VLMs and stimulate their respective strengths. These are not seriously discussed in previous works, and perhaps we should make some efforts in this regard.

David: Yes, we can try using LLM to read the output of multiple VLMs, then letting it reason about the best answer, and see whether LLM is a good coordinator in between VLMs.

[After a few days...]

David: Hi! Brian, I tried this paradigm on A-OKVQA dataset, I used BLIP and OFA as the base VLM, then combined their outputs into a prompt template, fed the prompt into FLAN, and then finetune FLAN for one epoch. This paradigm achieves results that far exceed the both single-model and ensemble performance using pretrained BLIP and OFA, and even outperforms the current state-of-the-art approach that requires much more epochs of finetuning.

Brian: No way! Largely surpassing single model performance may suggest that FLAN learns new knowledge after finetuning. This process may indicate FLAN is *coordinating* multiple VLMs. We could do more experiments to justify this conjecture.

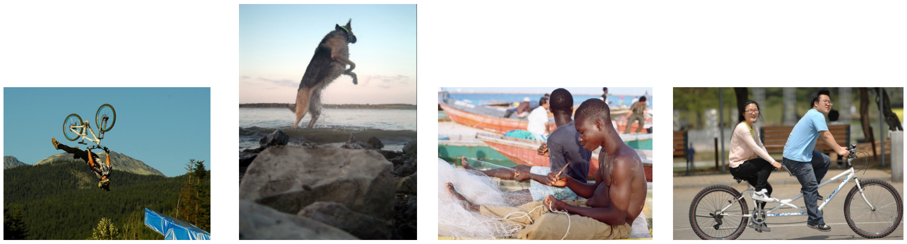
David: This may also be related to instruction tuning, FLAN is a language model that uses instructions to complete tasks, it may regard the prompt template as a kind of coordination instruction. And then follow the coordination instructions to integrate multiple VLMs outputs to conclude its final answer. Language models? coordination? That's interesting, we could probably call this paradigm Coordinative language models? Cola, a cool name.

Brian: This is incredibly fascinating! It's a great name, and this paradigm has so much potential. We should refine this study and tell the community about this finding!

[After it, the two students and their collaborators, the supervisors, began to focus their efforts studying on Cola.]

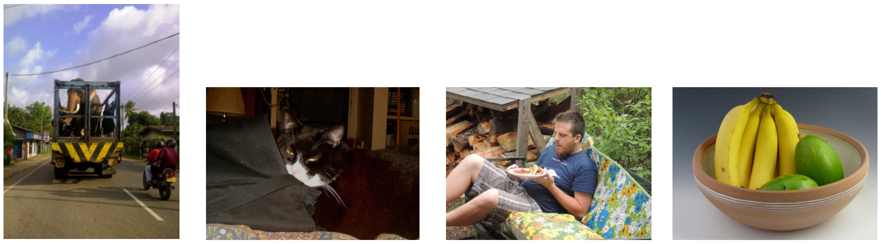
Question	Why might people sit here?	The room can be described as what?	In what type of location are they playing with the body board?	What is in front of the monitor?
OFA caption	colorful umbrellas on the riverwalk	living room layout and decor medium size how to decorate a small living room dining combo mant	person, left, and person look at a painting of a great white shark.	a desk with a computer, a lamp, a laptop, and a plant.
BLIP caption	a colorful umbrella umbrella with colorful umbrellas	a dining room table with a glass table and chairs	a man holding a surfboard while another man is standing next to him	a desk with a computer and a lamp
Choices	['to testify', 'to rest', 'to shop', 'get tattoo']	['tidy', 'messy', 'on fire', 'destroyed']	['room', 'beach', 'park', 'store']	['keyboard', 'phone', 'mouse', 'headphones']
OFA answer	to eat	living room	bedroom	a keyboard
BLIP answer	yes	dining room	beach	monitor
Cola-Zero answer	to rest	tidy	beach	keyboard
Cola-FT answer	to rest	tidy	room	keyboard

Figure 6: A-OKVQA qualitative examples. Leftmost: LM doesn't use BLIP and OFA's answers, but may observe from captions to derive the correct final answer. Left: As shown on the left, LM does not follow the wrong answers from OFA and BLIP but gets the correct answers from captions. Right: With both OFA and BLIP answering incorrectly, LM derives the correct one from both VLMs' captions and answers. Rightmost: After assessing the questions, answers, and captions, LM goes with OFA's answer and rewrites it to match the expression in choices. The correct choices are underlined. Cola-Zero answers are given in zero-shot settings.



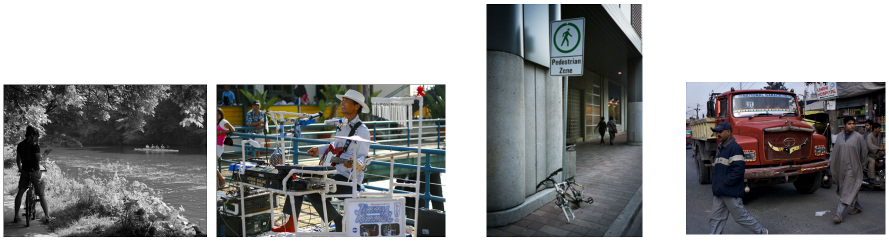
Question	Does the image describe " A professional daredevil "?	Does the image describe " the dog is a shitz " ?	Does this image describe "Two twenty-somethings prepare to catch salmon while other older men catch catfish" ?	Does this image describe "A little girl gets hit by a woman riding a bike" ?
OFA caption	person doing a flip on a mountain bike	a dog jumping out of the water.	men repairing fishing nets on the beach in zanzibar, tanzania	a man and a woman on a tandem bike
BLIP caption	a man doing a trick on a bike in the air	a dog jumping over rocks in the water	a man sitting on a boat with a fishing net net	a man and woman riding a bicycle in a parking lot
Choices	['yes', 'maybe', 'no']	['yes', 'maybe', 'no']	['yes', 'maybe', 'no']	['yes', 'maybe', 'no']
OFA answer	yes	no	yes	yes
BLIP answer	yes	no	yes	no
Cola-Zero answer	yes	no	no	no
Cola-FT answer	maybe	maybe	maybe	no

Figure 7: e-SNLI-VE qualitative examples. Leftmost: As the connection to *daredevil* is not obvious in BLIP and OFA's captions, although Cola-Zero is misled, Cola-FT correctly answers *maybe*. Left: Similar to the left example, Cola-FT answer correctly as no obvious connections are seen from the captions to this question. Right: Similar to the left example, the fact of *catch catfish* is not reasonable from the captions, Cola-FT picks the correct answer *maybe*. Rightmost: As *girl gets hit* is not obvious in BLIP and OFA's captions and answers, Cola-Zero and Cola-FT both follow BLIP to choose the correct answer *no*. The correct choices are underlined. Cola-Zero answers are given in zero-shot settings.



Question	Does this image describe "The truck contains the elephant" ?	Does this image describe "The bed is under the handbag" ?	Does this image describe "The couch is behind the hot dog" ?	Does this image describe "The bowl contains the banana" ?
OFA caption	an elephant being transported on a truck in sri lanka	a black and white tuxedo cat with a white nose, yellow eyes, and white	person enjoying a meal by the fire	bananas and mangoes in a bowl
BLIP caption	a truck with a large elephant in the back of it	a black cat laying on a bed with a pillow	a man sitting on a couch with a plate of food	a bowl of fruit is shown in this bowl
Choices	['yes', 'no']	['yes', 'no']	['yes', 'no']	['yes', 'no']
OFA answer	yes	no	yes	yes
BLIP answer	no	no	yes	no
Cola-Zero answer	no	no	no	yes
Cola-FT answer	yes	no	no	yes

Figure 8: VSR qualitative examples. Leftmost: As OFA caption mentioned *elephant being transported* and OFA provides the correct answer, Cola-FT follows OFA's choice. Left: As OFA and BLIP provide the same answer, Cola-Zero and Cola-FT follow the choice. Right: As the captions do not provide obvious information, even BLIP and OFA provide the same answer, Cola-Zero and Cola-FT are not misled to the wrong choice. Rightmost: As the captions provide strong clue *bananas in a bowl*, although BLIP's answer is incorrect, Cola-Zero and Cola-FT still choose the correct answer. The correct choices are underlined. Cola-Zero answers are given in zero-shot settings.



Question	What are the people doing in the water?	Does the image describe " The man is making a vase"?	What kind of zone is this bike parked in?	Does this image describe "The motorcycle is beside the truck" ?
OFA caption	black and white photo of a man on a bike looking at a canoe in the river	person on the potter's wheel	a city made by people bucharest	men walking past a truck in kabul, afghanistan.
BLIP caption	a man and woman on a bike in a park	a man is sitting on a chair and is using a wheel	a bicycle parked next to a pedestrian crossing sign	a man walking down the street in a city
Choices	['surfing', 'fishing', 'kayaking', 'swimming']	['yes', 'maybe', 'no']	['temporary', 'pedestrian', 'no parking', 'handicap']	['yes', 'no']
OFA answer	boating	yes	pedestrian	yes
BLIP answer	swimming	no	no parking	no
Cola-Zero answer	OFA	no	no parking	no
Cola-FT answer	kayaking	maybe	no parking	no

Figure 9: Failure cases. The correct choices are underlined. Cola-Zero answers are given in zero-shot settings.