Visual Question Answering with Annotation-Efficient Zero Shot Learning under Linguistic Domain Shift

Pratyay Banerjee, Tejas Gokhale, Yezhou Yang, Chitta Baral
Arizona State University
{pbanerj6, tgokhale, yz.yang, chitta}@asu.edu

Abstract

Heavy reliance on human-annotated training datasets (which typically suffer from annotator subjectivity and linguistic priors) has led to learning spurious correlations, bias amplification, and lack of robustness in vision-and-language (V&L) models. We study whether VQA models can be trained without any human-annotated Q-A pairs or object-bounding boxes. We use a self-supervised framework that involves procedural synthesis of Q-A pairs from captions and pre-training tasks for training our models. Since our Q-A pairs are synthetic, they exhibit a linguistic domain shift from the questions in VQA data and a label-shift in the answer-set, i.e. a zero-shot learning task. We benchmark our models on VQA-v2, GQA, and on VQA-CP which contains a softer version of label shift.

1 Introduction

Visual question answering (VQA) has emerged as a crucial task in visual understanding, and human-annotated datasets [1, 2, 3, 4, 5] have been used to train and evaluate various VQA models. Unfortunately, heavy reliance on these datasets for training has resulted in introduction of bias towards answer styles, question-types [6], and spurious correlations with language priors [7, 8, 9, 10]. As such, evaluation of VQA models on test-sets very similar in style to the training samples, is deceptive and inadequate, and not a true measure of robustness. Work in VQA has recently attracted attention under points of view of robustness, reduction of biases and spurious correlations. Performance under **domain shift** has been evaluated for test questions with unseen words [11], unseen objects [12], novel compositions [13, 14], logical connectives [15], varying linguistic styles [6, 16, 17, 18] and different reasoning capabilities [19, 20, 21]. **Label shift** has been implicitly hinted at in VQA-CP [7]. To mitigate such robustness challenges, one line of work has focused on balancing, de-biasing, and diversifying samples [22, 23]. However crowdsourcing "unbiased" labels is costly, and requires a well-designed annotation interface, large-scale human effort and time [24]. The alternative is to avoid the use of annotations, and instead train models by synthesizing training data [25].

In this work, we train VQA models without using human-annotated Q-A pairs. We utilize image-captioning datasets which provide a multi-perspective concise description of visible objects in an image, and procedurally generate Q-A pairs using a self-supervised mechanism. Since our Q-A pairs are created synthetically, there exists a domain shift as well as label (answer) shift from evaluation datasets as shown in Figure 1, making it a zero-shot learning task. We propose pre-training tasks that use spatial pyramids of image-patches instead of object bounding-boxes, further making our method label-efficient, and removing the dependence on labeled object bounding boxes. We extensively evaluate two models, UpDown [26] and a transformer-encoder [27] based model pre-trained on synthetic Q-A pairs and image-caption matching task, and analyze them under zero-shot and fully-supervised settings, to establish benchmarks on VQA-v2, VQA-CP, and GQA. Our model serves as a strong baseline for future work on zero-shot VQA.

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

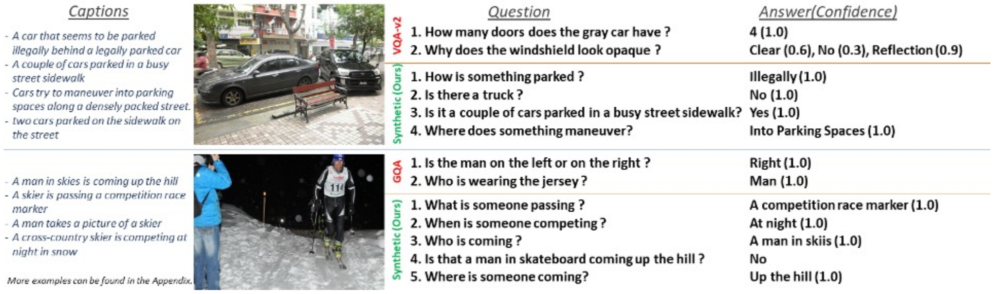


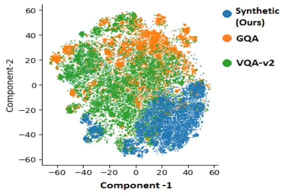
Figure 1: Images from VQA and GQA with human-annotated (red) and synthetic (green) Q-A pairs.

2 Self-Supervised Data Synthesis and Pre-Training Framework

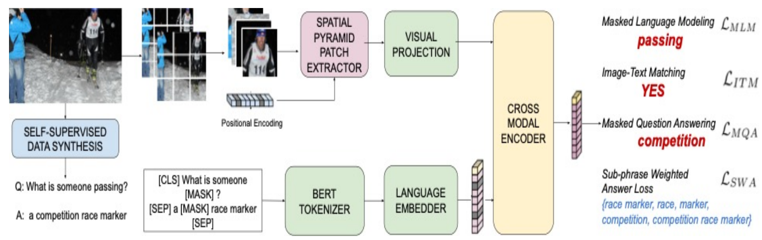
Why Captions? Image captioning is a crucial vision-and-language task, and datasets such as MS-COCO [28] contain captions that describe common objects and actions. During the construction of MS-COCO, human caption writers were instructed to refrain from describing the past or future or "what a person might say". One the other hand, annotators of VQA [2] were instructed to ask "interesting" questions that may require "commonsense" and could fool a robot, and were allowed to *speculate* an answer. In Figure 1, the first VQA question demonstrates linguistic bias since most cars have four doors, the second question is subjective and receives contradicting answers from annotators, while the first GQA question is ambiguous and could refer to either the skier or the photographer. Thus the nature of the data-collection design for VQA introduces human subjectivity and linguistic bias, as compared to image captions. Motivated by this, we show that procedural generation of Q-A pairs from captions can lead to a diverse variety of questions that need deep visual understanding.

2.1 Question-Answer Data Synthesis:

In this section, we detail our framework for generating Q-A pairs from captions. Questions are categorized based on their answer types; *Yes-No*, *Number*, *Color*, *Location*, *Object* and *Phrases*.
Template-Based: To create *Yes-No* questions, modal verbs are removed from the caption, and a randomly chosen question prefix such as *"is there", "is this", "does this look like"* is attached, for e.g. "A man is wearing a hat and sitting" → ("*Is there a* man wearing a hat and sitting", "Yes"). For *Object*, *Number*, *Location*, and *Color* questions, we follow a procedure similar to [29]. To create "what" questions for the *Object* type, we replace objects with "what", and rephrase the question. Similarly for *Number* questions; we extract numeric quantifiers of noun phrases, and ask "how many" and "what is the count" questions. *Color* questions are generated by locating the a noun-phrase and its color-adjective, and replacing them in a templated question: "What is the color of the object?". *Location* questions are similar to *Object* questions, but we extract phrases with "in", "within" to extract locations, with places, scenes, and containers as answers.
Semantic Role Labeling: QA-SRL [30] was proposed as an annotation paradigm that uses Q-A pairs to specify textual arguments and their roles. For the caption *"A girl in a red shirt holding a skateboard sitting in an empty open field"*, using QA-SRL with B-I-O span detection and sequence-to-sequence models [31], we obtain *"when", "what", "where", and "who"* questions, belonging to the *Phrases* category such as: *(what is someone holding?, a skateboard), (who is sitting?, a girl in a red shirt holding a skateboard), (where is someone sitting?, an empty open field)* QA-SRL questions are short and use generic descriptors and pronouns such as *something* and *someone* instead of elaborate references, while the expected answer phrases are longer and descriptive.
Paraphrasing and Back-Translation (P&B): To paraphrase questions, we train a T5 [32] text generation model on the Quora Question Pairs Corpus [33]. For back-translation we train another T5 text generation model on the Opus corpus [34], translate the question to an intermediate language (Français, Deutsche, or Español) and re-translate the question back to English.
Comparitive Analysis with VQA and GQA: QA-SRL questions require semantic understanding of the actions depicted in the image, and answers to these are more descriptive with use of adjectives, adverbs, determiners, and quantifiers, compared to current VQA benchmarks, which typically contain one-word answers. Our synthetic data contains 90k unique answer phrases, compared to 3.2k in VQA and 3k in GQA. Figure 1 shows significant overlap between two human-annotated datasets VQA and GQA, while our synthetic questions display a domain shift.



(a) t-SNE plot showing the discrepancy between VQA, GQA and our synthetic samples.



(b) Overview of our self-supervised pre-training framework and model.

2.2 Method

Recent approaches [35, 36, 37] use transformer encoders pre-trained on V&L tasks on a combination of multiple captioning and VQA datasets [38, 39, 40, 28], which is resource-intensive. Instead, we train our models only on less noisy and multi-perspective image descriptions from MS-COCO.
Spatial Pyramid Patches: "Bottom-Up" object region features [26] have become the de-facto image features used in VQA models, but do not represent non-object regions and backgrounds which may be necessary for VQA. This is a problematic bottle-neck, since object detectors can be incorrect, and can fail to detect rare and small objects [41]. Inspired by SPM [42] for image classification, we propose *spatial pyramid patch features* to represent the input image into a multi-scale sequence of features. We use a ResNet (pretrained on ImageNet) to extract features from a grid of image patches. Larger patches encode global features and relations, while smaller patches encode local features.
Encoder: Our Encoder model is similar to the UNITER single-stream transformer, where the sequence of word tokens $w = \{w_1, ..., w_T\}$ and the sequence of image patch features $v = \{v_1, ..., v_K\}$ are taken as input. The visual features are projected to a shared embedding space using a fully-connected layer. A projected visual position encoding, indicating the patch region is concatenate with the visual features and used as input to L cross-modality attention layers.

We train the Encoder model using three pre-training tasks listed below:
Masked Language Modeling (MLM) We randomly mask 15% of the word tokens from the caption and train the model to predict them, as in [35]. For instance, when the model receives the input "There is [MASK] wearing a hat", without the image, there can be multiple plausible choices, such as "woman", "man", "girl". But given the image, the model should predict "man".
Masked Question Answering (MQA): In this task, the answer tokens are masked, and the model is trained to predict the answer tokens. For example in Figure 1, for the input " When is someone competing? [MASK] [MASK]", the model should predict, "at night".
Image-Text Matching (ITM): For each image, we use the five MS-COCO captions as positive samples, and randomly captions from other images with different set of objects as negative samples. We train the model on a binary matching task for each image-caption pair.
Sub-phrase Weighted Answer Loss: We extract all possible sub-phrases that can be alternate answers, but assign them a lower weight than the complete phrase. We train the model with an additional sub-phrase weighted loss ($\mathcal{L}_{SWA} = \mathcal{L}_{BCE}(\sigma(z^{[CLS]}), y_{wa})$), which enforces a distribution over the probable answer vocabulary y_{wa} instead of a single true answer.

3 Experiments and Results

Datasets and Baselines: We evaluate our methods on three popular benchmarks: VQA-v2, VQA-CP v2, and GQA, under the *zero-shot setting* (trained only on procedurally generated samples), and with *fully-supervised finetuning* of our model on human-annotated samples. We measure the improvements due to our proposed image patch features and SWA loss, as compared to UpDn [26], which uses object bounding-box features. Since pre-trained transformers [37, 35] use large and densely annotated V&L corpora for supervision, we compare with these only in the fully-supervised setting.

Zero-shot Question Answering (ZSL): Tables 13, 14 and 15 summarize our results on the three benchmark datasets respectively. Our method (with the Encoder model) outperforms specially designed supervised methods for bias removal in VQA-CP, with model-agnostic performance improvements for both UpDn and Encoder models. In the ZSL setting, compared to object-features, our annotation-efficient method of Spatial Image Patch features are better on VQA-CP and competitive

Table 1: VQA-CP-v2 test.

Model	All	YN	Num	Oth
SAN [45]	25.0	38.4	11.1	21.7
GVQA [7]	31.3	58.0	13.7	22.1
UpDn [26]	39.1	62.4	15.1	34.5
AReg[12]	42.0	65.5	15.9	36.6
AdvReg [46]	42.3	59.7	14.8	40.8
RUBi [47]	47.1	68.7	20.3	43.2
[48]	46.0	58.2	29.5	44.3
Unshuffling [49]	42.4	47.7	14.4	47.3
UpDn+CE+GS [50]	46.8	64.5	15.4	45.9
LXMERT [35]	46.2	42.8	18.9	**55.5**
ZSL+Objects+UpDn	40.8	67.4	28.6	30.2
ZSL+Patches+UpDn	41.2	68.5	29.8	30.0
ZSL+Patches+Enc	**47.3**	**73.4**	**39.8**	35.6

Table 2: VQA-v2 Test-standard.

Model	All	YN	Num	Oth
GVQA [7]	48.2	72.0	31.1	34.7
UpDn [26]	65.3	81.8	44.2	56.1
RUBi [47]	63.1	*	*	*
MCAN [51]	70.4	85.8	53.7	60.7
VilBERT [36]	70.5	*	*	*
LXMERT [35]	72.5	**88.2**	**54.2**	**63.1**
UNITER [37]	**72.7**	*	*	*
ZSL + Objects + UpDn	41.4	68.1	27.6	29.4
ZSL + Patches + UpDn	40.6	67.8	28.4	29.2
ZSL + Patches + Enc	46.8	72.1	34.4	34.1
FSL + Patches + UpDn	63.4	80.2	45.2	52.1
FSL + Patches + Enc	65.3	80.5	48.94	56.2

Table 3: GQA Validation.

Model	All	Binary	Open
CNN + LSTM [43]	46.6	61.9	22.7
UpDn [26]	49.7	66.6	34.8
MAC [43]	54.1	71.2	38.9
BAN [44]	57.1	76.0	40.4
LXMERT [35]	**60.3**	**77.8**	**45.0**
ZSL + Objects + UpDn	30.7	50.8	17.6
ZSL + Patches + UpDn	31.1	52.3	16.8
ZSL + Patches + Enc	33.7	55.5	21.2
FSL + Patches + UpDn	46.4	64.3	31.4
FSL + Patches + Enc	55.2	73.6	38.8

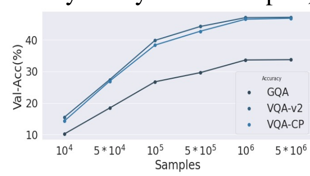
Table 4: Impact of data-synthesis methods on ZSL performance.

	Datasets	Temp	+P&B	QASRL	All
Updn	VQA-v2	26.2	28.5	31.1	41.4
	VQA-CP	25.7	27.1	33.8	40.2
	GQA	11.6	14.8	18.9	31.1
Encoder	VQA-v2	32.5	34.8	40.3	**47.1**
	VQA-CP	31.2	33.6	39.8	**46.8**
	GQA	18.5	23.6	21.4	**33.7**

Table 5: Effect of pretraining tasks on Encoder ZSL performance.

Datasets	SWA	MLM +SWA	MQA +SWA	MLM +MQA +SWA	MLM +IT +SWA	All
VQAv2	39.1	42.4	42.0	45.6	44.7	**46.2**
VQACP	38.3	41.5	41.2	44.9	43.6	**45.4**
GQA	25.4	27.8	26.6	29.7	28.9	**31.2**

Table 6: Learning Curve (accuracy vs #synthetic samples)



on VQA. ZSL performance for GQA is not as effective, which could be attributed to the lack of questions about spatial relationships in our synthetic samples, which are crucial for the GQA task.
Fully Supervised Question Answering (FSL): The performance of our methods when finetuned on human-labeled samples approaches SOTA methods. In GQA, the Encoder model performs on par with MAC [43] and BAN [44] that unlike us use object-relation labels, suggesting that pyramidal features are capable of learning spatial relationships between image regions.

Ablation Studies: We perform analyses and ablation studies to establish the efficacy various components of our method. *Details and insights from these can be found in the Appendix.*
Table 16 shows the effect of different question generation techniques with the largest gains due to QA-SRL based questions and the SWA-Loss. Table 18 shows the effect of different pretraining tasks on the downstream zero-shot VQA task. The combination of MLM, MQA and ITM, all of which need image understanding, shows improved performance on the downstream task, indicating better cross-modal representations. Figure 6 shows the learning curve of our Encoder model for the zeroshot setting trained on our synthetic Q-A pairs. The performance stagnates after a critical threshold of 10^6 samples is reached. Our experiments also suggest that randomly sampling a set of questions for each image per epoch leads to a $+4\%$ gain, as compared to training on the entire set.

Error Analysis As our ZSL method is pretrained on longer phrases, it tends to generate answers with more details, such as “red car" instead of “car". The SWA loss mitigates this to an extent but the bias towards short answers is not completely removed. We observe that for 42% of questions the target answer is a sub-phrase of our predicted answer, and for 87% of such samples, our detailed predictions are indeed plausible answers. This shows the utility in learning from captions, and also quantifies the bias towards short “true" human-labeled answers, thus demonstrating the need for better evaluation metrics that do not penalize descriptive answers. In the fully supervised setting, the pre-trained QA classifier continues to predict longer phrases as answers, leading to a drop in accuracy, while the feedforward layer (trained on human annotations) performs better ($+6\%$), indicating our Encoder captures relevant features necessary to generalize to the benchmark answer-space.

4 Discussion and Conclusion

Prior work [26, 35, 37] has resulted in notable improvements for V&L tasks using object bounding-boxes and region features. But there is little effort towards developing equally reliable methods that do not depend upon dense annotations. In this work, we seek a pathway for the V&L community towards annotation-efficiency via self-supervised techniques. We present a framework for procedural synthesis of Q-A pairs and introduce the new task of zero-shot VQA, where benchmark datasets can be used *only* for evaluation. We demonstrate the potential of replacing object-features with annotation-efficient spatial pyramids of patch features. Our method surpasses previous supervised backbone methods on VQA-CP.

References

[1] Mateusz Malinowski and Mario Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In *Advances in neural information processing systems*, pages 1682–1690, 2014.

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

[3] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *CVPR*, 2018.

[4] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, pages 6700–6709, 2019.

[5] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *CVPR*, June 2019.

[6] Wei-Lun Chao, Hexiang Hu, and Fei Sha. Cross-dataset adaptation for visual question answering. In *CVPR*, pages 5716–5725, 2018.

[7] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *CVPR*, 2018.

[8] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. Annotation artifacts in natural language inference data. In *NAACL-HLT (2)*, 2018.

[9] Timothy Niven and Hung-Yu Kao. Probing neural network comprehension of natural language arguments. In *57th Annual Meeting of the ACL*, pages 4658–4664, 2019.

[10] Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*, 2019.

[11] Damien Teney and Anton van den Hengel. Zero-shot visual question answering. *arXiv preprint arXiv:1611.05546*, 2016.

[12] Santhosh K Ramakrishnan, Ambar Pal, Gaurav Sharma, and Anurag Mittal. An empirical evaluation of visual question answering for novel objects. In *CVPR*, pages 4392–4401, 2017.

[13] 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 *CVPR*, 2017.

[14] Aishwarya Agrawal, Aniruddha Kembhavi, Dhruv Batra, and Devi Parikh. C-vqa: A compositional split of the visual question answering (vqa) v1.0 dataset. *arXiv preprint arXiv:1704.08243*, 2017.

[15] Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. Vqa-lol: Visual question answering under the lens of logic. In *European Conference on Computer Vision (ECCV)*, 2020.

[16] Yiming Xu, Lin Chen, Zhongwei Cheng, Lixin Duan, and Jiebo Luo. Open-ended visual question answering by multi-modal domain adaptation. *arXiv preprint arXiv:1911.04058*, 2019.

[17] Robik Shrestha, Kushal Kafle, and Christopher Kanan. Answer them all! toward universal visual question answering models. In *CVPR*, pages 10472–10481, 2019.

[18] Ramprasaath R Selvaraju, Purva Tendulkar, Devi Parikh, Eric Horvitz, Marco Tulio Ribeiro, Besmira Nushi, and Ece Kamar. Squinting at vqa models: Introspecting vqa models with sub-questions. In *CVPR*, pages 10003–10011, 2020.

[19] Kushal Kafle and Christopher Kanan. An analysis of visual question answering algorithms. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1965–1973, 2017.

[20] Marco Tulio Ribeiro, Carlos Guestrin, and Sameer Singh. Are red roses red? evaluating consistency of question-answering models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6174–6184, July 2019.

[21] Arijit Ray, Karan Sikka, Ajay Divakaran, Stefan Lee, and Giedrius Burachas. Sunny and dark outside?! improving answer consistency in vqa through entailed question generation. In *Proceedings of the 2019 Conference on EMNLP*, 2019.

[22] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, pages 6904–6913, 2017.

[23] Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Yin and yang: Balancing and answering binary visual questions. In *CVPR*, 2016.

[24] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. In *AAAI*, 2020.

[25] Patrick Lewis, Ludovic Denoyer, and Sebastian Riedel. Unsupervised question answering by cloze translation. In *Proceedings of the 57th Annual Meeting of the ACL*, pages 4896–4910, 2019.

[26] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, June 2018.

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[28] 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*, pages 740–755. Springer, 2014.

[29] Mengye Ren, Ryan Kiros, and Richard Zemel. Exploring models and data for image question answering. In *Advances in neural information processing systems*, pages 2953–2961, 2015.

[30] Luheng He, Mike Lewis, and Luke Zettlemoyer. Question-answer driven semantic role labeling: Using natural language to annotate natural language. In *Proceedings of 2015 the conference on EMNLP*, pages 643–653, 2015.

[31] Nicholas FitzGerald, Julian Michael, Luheng He, and Luke Zettlemoyer. Large-scale qa-srl parsing. In *56th Annual Meeting of the ACL*, 2018.

[32] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019.

[33] Shankar Iyer, Nikhil Dandekar, and Kornel Csernai. First quora dataset release: Question pairs, 2017.

[34] Jorg Tiedemann. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, 2012.

[35] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *EMNLP 2019*, 2019.

[36] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23, 2019.

[37] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Learning universal image-text representations. *arXiv preprint arXiv:1909.11740*, 2019.

[38] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the ACL*, pages 2556–2565, 2018.

[39] Vicente Ordonez, Girish Kulkarni, and Tamara L Berg. Im2text: Describing images using 1 million captioned photographs. In *Advances in neural information processing systems*, 2011.

[40] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.

[41] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Meta-learning to detect rare objects. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9925–9934, 2019.

[42] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[43] Drew A Hudson and Christopher D Manning. Compositional attention networks for machine reasoning. In *International Conference on Learning Representations*, 2018.

[44] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. In *Advances in Neural Information Processing Systems*, pages 1564–1574, 2018.

[45] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *CVPR*, pages 21–29, 2016.

[46] Gabriel Grand and Yonatan Belinkov. Adversarial regularization for visual question answering: Strengths, shortcomings, and side effects. In *Proceedings of the Second Workshop on Shortcomings in Vision and Language*, pages 1–13, 2019.

[47] Remi Cadene, Corentin Dancette, Hedi Ben younes, Matthieu Cord, and Devi Parikh. Rubi: Reducing unimodal biases for visual question answering. In *Advances in Neural Information Processing Systems 32*, pages 841–852. 2019.

[48] Damien Teney and Anton van den Hengel. Actively seeking and learning from live data. In *CVPR*, pages 1940–1949, 2019.

[49] Damien Teney, Ehsan Abbasnejad, and Anton van den Hengel. Unshuffling data for improved generalization. *arXiv preprint arXiv:2002.11894*, 2020.

[50] Damien Teney, Ehsan Abbasnedjad, and Anton van den Hengel. Learning what makes a difference from counterfactual examples and gradient supervision. *arXiv preprint arXiv:2004.09034*, 2020.

[51] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *CVPR*, 2019.

[52] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on EMNLP*, pages 1532–1543, 2014.

[53] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771, 2019.

[54] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d Alche-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

Appendix

A Synthesized Samples

In this section, we will show illustrative examples of our procedurally generated question-answer (Q-A) pairs. Table 7 shows examples of questions and answers generated from the image caption using our template-based method. Corresponding images are shown for clarity. Table 8 shows the use of two transformations (T): negation and adversarial words [15] two generate more sentences. Thus the negation of Q or substitution of a word in Q with an adversarial word results in the new question-answer pair Q_{new}, A_{new}. To increase the linguistic diversity of the questions we use paraphrasing as shown in Table 9.

Table 7: Examples of template-based generation of our self-supervised data synthesis framework

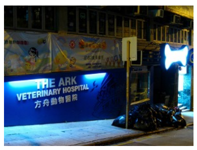
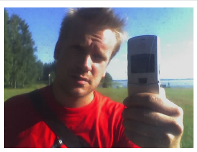
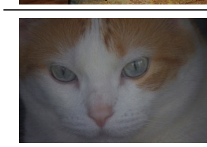
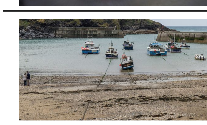
Image	Question	Answer
	What are set on the sidewalk outside a veterinary hospital?	bags
	What is the young man holding up in front of his face ?	phone
	What is almost empty on the table	glass
	What drawn carriage with passengers in the city	horse
	What is the color of the table ?	white
	What is the color of the eyes ?	blue
	How many boats anchored by ropes close to shore?	8

Table 8: The effect of using transformations (T) to create new Q-A pairs

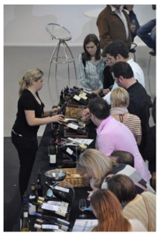
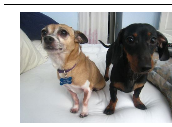
T	Image	Q	A	$\mathbf{Q_{new}}$	$\mathbf{A_{new}}$
Negation		Is this bread?	yes	Is this not bread	no
		What is the color of the woman's shirt?	black	What is not the color of the woman's shirt?	white
		Is there a boy?	no	Is there no boy?	yes
Adversarial		Who is sitting in the boat ?	man	Who is sitting in the dining table ?	"can't say"
		How big is the plane ?	large	How big is the car ?	"size"
		How many puppies are on the bed ?	two	How many cats are on the bed?	none

Table 9: Illustration of using paraphrasing to improve the linguistic variation of our questions and answers.

Image	Q	A	$\mathbf{Q_{new}}$	$\mathbf{A_{new}}$
	How is something parked ?	illegally	How's-what's parked?	illegally
	what does something seem to do ?	park	What do you think something seems to be doing?	park
	Where was parked something?	behind a legally parked car	Do you know where something was parked?	behind a legally parked car
	How many cars are visible ?	2	How many cars are we looking at?	2
	Is there two cars parked on the sidewalk on the street ?	Yes	There are two cars parked on the sidewalk, right?	Yes

B Dataset Analysis

Answers to QA-SRL questions are more descriptive with use of adjectives, adverbs, determiners, and quantifiers, compared to current VQA benchmarks, which typically contain one-word answers. Similarly, questions have less descriptive subjects due to the use of pronouns. Dataset statistics comparing our data with benchmarks datasets are shown in Table 10.

We also observe there are around 200 answers that are not present in our answer phrases, such as time (11:00) and proper nouns (LA Clippers), both of which are not present in caption descriptions. The style of some of our synthetic questions such as counting questions, object presence/absence questions created by template-based question generation, is also found in VQA and GQA. On the

Table 10: Dataset statistics for our generated Q-A pairs. Train/Validation sample counts for benchmark datasets are provided.

	Template-based	Paraphrase & Back-translate	QA-SRL	VQA-v2	GQA	VQA-CP
# of Questions	600K	400K	2.5M	438K / 214K	943K / 132K	245K / 220K
# of Answers	5K	5K	90K	3.5K	1878	3.5K
Mean Question Length	7.9	8.1	4.8	6.4	10.6	6.4
Mean Answer Length	1.4	1.4	6.3	1.1	1.3	1.1
Image Source	COCO	COCO	COCO	COCO	COCO,Genome,Flickr	COCO
Image Counts	204K	204K	204K	204K	113K	204K

Table 11: Distribution of samples by answer-type in our pre-training dataset and the VQA-CP dataset used for evaluation.

Category	VQA-CP (%)	Pretraining (%)
Yes/No	41.86	50.18
Number	11.91	8.32
Other	46.23	41.45

other hand, QA-SRL questions require semantic understanding of the actions (verb) depicted in the image, which are rare in VQA and GQA.

We compare the distribution per answer-type of our synthetically generated samples with the distribution in the VQA-CP-v2 [7] dataset ub Table 11. Since we use our synthetic samples as the pre-training data, and do not use VQA-CP samples for training in our zero-shot setup, this comparison shows us that there is a shift between the training (synthetic) and test (human annotated VQA-CP) datasets.

We further analyze this shift, by computing the t-SNE projections of questions using mean-pooled Glove [52] embeddings for our generated questions and observe the overlap with human-authored questions in VQA and GQA [4]. Figure 3. We observe a marked shift between the question clusters for our procedurally generated questions and human annotated questions from VQA and GQA.

Similarly, we also show the distribution of answers in our dataset in Figure 4. It can be seen that our dataset has a slight imbalance in the proportion of questions with answer "yes" and "no". Numeric answers *0,1,2,3* are most frequent. Answers about people such as *man, woman, people, person, group of people* are also more common in the dataset. The remaining answers have a long-tailed distribution, since there are $\sim 90k$ unique answers in our dataset compared to $\sim 3.5k$ in VQA and $\sim 2k$ in GQA.

C Experimental Setup

C.1 Datasets

We evaluate our methods on the three popular visual question answering benchmarks: VQA-v2, VQA-CP v2, and GQA. Answering questions in VQA-v2 and VQA-CP v2 requires image and question understanding, whereas GQA further requires spatial understanding such as compositionality and relations between objects. We evaluate our methods under *zero-shot* (trained only on procedurally generated samples), and *fully-supervised* (where we finetune our model using the associated train annotations) settings. We report exact-match accuracies as our metrics for evaluation.

C.2 Training

Our Encoder has 8 cross-modal layers with a hidden dimension of 768. Our models are pre-trained for 40 epochs with a learning rate of $1e{-5}$, batch size of 256, using Adam optimizer. For finetuning, we use a learning rate of $1e{-5}$ or $5e{-5}$ and batch size of 32 for 10 epochs. We use a ResNet-50 pretrained on ImageNet to extract features from image patches with 50% overlap, and Faster R-CNN pretrained on Visual Genome to extract object features. We use the HuggingFace [53] and Pytorch Deep learning framework [54]. Hyperparameters and other training settings are given in Table 12 All our models are trained using 4 Nvidia V100 16 GB GPUs. Our code will be made available upon publication.

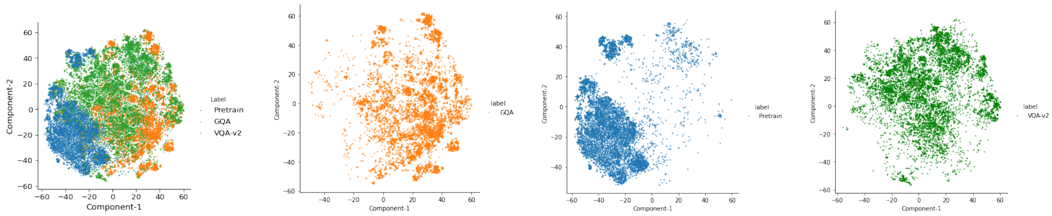
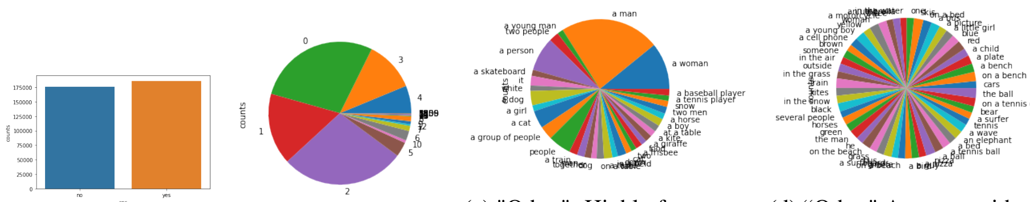


Figure 3: t-SNE projections of Glove embedding our generated questions, and human-authored VQA-v2 and GQA questions. Blue: our pretraining dataset, Orange: GQA, Green: VQA. L-R: All, GQA, Pretrain, VQA.



(a) “Yes-No" (b) “Numeric" (c) "Other": Highly frequent answers (count > 500) (d) “Other":Answers with count between 200 and 500

Figure 4: Distribution of most frequent answers in our Pretraining dataset for each answer-type (yes-no, numeric, and other). Please zoom for details.

Table 12: Hyper-Parameters for our models

Hyper-Parameters	**Model**
Batch Size	32-128
Learning Rate	$1e^{-5}$-$5e^{-5}$
Dropout	0.1
Language Layers	6
Cross-Modality Layer	4-12
Optimizer	BertAdam
Warmup	0.1
Max Gradient Norm	5.0
Max Text Length	30
ResNet	50,101,152
Epochs	10-40

D Results

In this section, we discuss our results and outcomes from analyses. ZSL refers to zero-shot setting and FSL refers to our models further finetuned on the respective train split.

D.1 Zero-shot Question Answering

Tables 13, 14 and 15 summarize our results on the three benchmark datasets respectively. We can observe that our method outperforms specially designed supervised methodsfor bias removal in VQA-CP. Our procedurally generated Q-A pairs improve performance for both UpDown and Encoder models, showing the method to be effective, and that the improvements are model-agnostic. Our Encoder model further improves the performance by 5.5% over the UpDown baseline. In the zero-shot setting, compared to object-features, our Spatial Image Patch features perform equally well on VQA, and are better on VQA-CP, and are also more annotation efficient. In GQA, the zero-shot performance is not as competetive when compared to our performance on VQA and VQA-CP. We attribute this to the need for understanding spatial relationships answer GQA questions. Such questions are infrequent in our synthetic training data since human-annotated captions do not contain

Table 13: Unsupervised accuracy on VQA-CP-v2 test set. All baselines are supervised methods trained on the train split. [2]

Model	All	Yes-No	Num	Others
SAN [45]	25.0	38.4	11.1	21.7
GVQA [7]	31.3	58.0	13.7	22.1
UpDown [26]	39.1	62.4	15.1	34.5
AReg[12]	42.0	65.5	15.9	36.6
AdvReg [46]	42.3	59.7	14.8	40.8
RUBi [47]	47.1	68.7	20.3	43.2
[48]	46.0	58.2	29.5	44.3
Unshuffling [49]	42.4	47.7	14.4	47.3
UpDn+CE+GS [50]	46.8	64.5	15.4	45.9
LXMERT [35]	46.2	42.8	18.9	**55.5**
ZSL+Objects+UpDown	40.8	67.4	28.6	30.2
ZSL+Patches+UpDown	41.2	68.5	29.8	30.0
ZSL+Patches+Encoder	<u>**47.3**</u>	<u>**73.4**</u>	<u>**39.8**</u>	<u>35.6</u>

Table 14: VQA-v2 Test-standard accuracies[2]. FSL models are pretrained on synthetic samples, and further finetuned on VQA-v2 train split. *not available

Model	All	Yes-No	Num	Others
GVQA [7]	48.2	72.0	31.1	34.7
UpDown [26]	65.3	81.8	44.2	56.1
RUBi [47]	63.1	*	*	*
MCAN [51]	70.4	85.8	53.7	60.7
VilBERT [36]	70.5	*	*	*
LXMERT [35]	72.5	**88.2**	**54.2**	**63.1**
UNITER [37]	**72.7**	*	*	*
ZSL + Objects + UpDown	41.4	68.1	27.6	29.4
ZSL + Patches + UpDown	40.6	67.8	28.4	29.2
ZSL + Patches + Encoder	<u>46.8</u>	<u>72.1</u>	<u>34.4</u>	<u>34.1</u>
FSL + Patches + UpDown	63.4	80.2	45.2	52.1
FSL + Patches + Encoder	65.3	80.5	48.94	56.2

detailed spatial relationships among objects. The development of self-supervised techniques to perform spatial reasoning is an interesting future direction for research.

D.2 Fully Supervised Question Answering

In the fully supervised setting, the performance of our methods approaches SOTA methods. However, our methods are significantly annotation-efficient as we only adopt COCO captions without dense object annotations during pre-training or training. In GQA, the Encoder model performs on par with MAC [43] and BAN [44], which unlike us, use object relationship annotations. This suggests that pyramidal features and the cross-modal transformer encoder layers can learn spatial relationships between image regions.

D.3 Impact of each question-generation technique

In Table 16 we can observe the effect of different question generation techniques. All models use spatial image patch features. QA-SRL based questions and the SWA-Loss contribute the most towards gains in performance, and the paraphrased questions provide larger linguistic variation.

[2]In all tables <u>underline</u> implies unsupervised best, and **bold** implies overall best. Baselines are trained on VQA/VQA-CP/GQA training data and our models on synthetic self-supervised data.

Table 15: GQA Validation split accuracies.[2]

Model	All	Binary	Open
CNN + LSTM [43]	46.6	61.9	22.7
UpDown [26]	49.7	66.6	34.8
MAC [43]	54.1	71.2	38.9
BAN [44]	57.1	76.0	40.4
LXMERT [35]	**60.3**	**77.8**	**45.0**
ZSL + Objects + UpDown	30.7	50.8	17.6
ZSL + Patches + UpDown	31.1	52.3	16.8
ZSL + Patches + Encoder	<u>33.7</u>	<u>55.5</u>	<u>21.2</u>
FSL + Patches + UpDown	46.4	64.3	31.4
FSL + Patches + Encoder	55.2	73.6	38.8

Table 16: Effect of different training data sources on ZSL validation accuracy. P&B Paraphrasing and Back-translation.

	Datasets	Template	Template + P & B	QASRL	All
Updn	VQA-v2	26.2	28.5	31.1	41.4
	VQA-CP	25.7	27.1	33.8	40.2
	GQA	11.6	14.8	18.9	31.1
Encoder	VQA-v2	32.5	34.8	40.3	**47.1**
	VQA-CP	31.2	33.6	39.8	**46.8**
	GQA	18.5	23.6	21.4	**33.7**

Table 18: Effect of different Pre-training tasks on the ZSL validation accuracies for the Encoder model.

Datasets	SWA	MLM+ SWA	MQA+ SWA	MLM+MQA +SWA	MLM+IT SWA	All
VQA-v2	39.1	42.4	42.0	45.6	44.7	**46.2**
VQA-CP	38.3	41.5	41.2	44.9	43.6	**45.4**
GQA	25.4	27.8	26.6	29.7	28.9	**31.2**

D.4 Effect of Spatial Pyramids

We study the effect of progressively increasing the number of overlapping spatial image patches (i.e. decreasing the patch size). It can be observed in Table 17 that an optima exists at grid-size of 7×7 after which the addition of smaller patches is detrimental. Similarly, only using patches of large size does not allow models to focus on specific regions of the image. Thus a trade-off exists between global context and region-specific features. We observe a minor improvement of 0.01-0.3% by extracting features from ResNet-101 compared to ResNet-50. Removing visual position embeddings has a significant effect on performance, with a drop of 4.6% to 8% on average, in both ZSL and FSL settings for VQA and GQA.

D.5 Effect of different Pre-training Tasks

Table 18 shows the effect of different pretraining tasks on the downstream zero-shot VQA task. We need the SWA task, as it is used to perform the zeroshot QA task. The combination of MLM, MQA and ITM, all of which need image understanding, shows improved performance on the downstream task, indicating better cross-modal representations.

Table 17: Effect of the number of spatial patches on ZSL validation accuracies with UpDn and our Encoder. {3,5} implies division of the image into a 3x3 and 5x5 grid of patches.

	Datasets	{1}	{1,3}	{1,3,5}	{1,3,5,7}	{1,3,5,7,9}
UpDn	VQA-v2	18.8	36.7	40.1	**41.4**	39.8
	VQA-CP	19.7	35.9	39.7	**40.2**	38.4
	GQA	11.3	24.5	29.5	**31.1**	29.3
Encoder	VQA-v2	26.4	42.6	44.3	**47.1**	46.2
	VQA-CP	27.7	43.1	45.2	**46.8**	45.4
	GQA	15.3	28.8	30.9	**33.7**	31.2

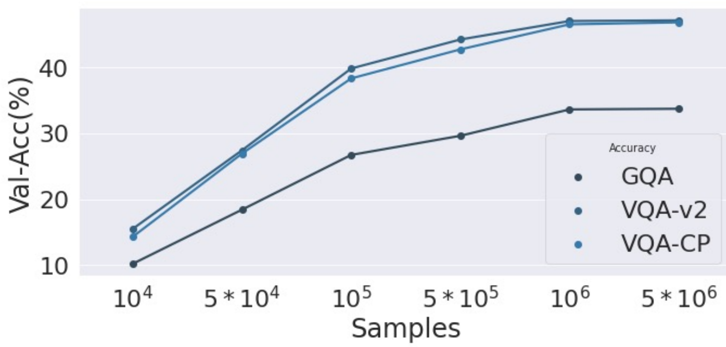


Figure 5: Learning Curve showing validation accuracy vs. the number of synthetically generated training samples.

D.6 Effect of size of Synthetic Train set

Figure 5 shows the learning curve of our Encoder model for the zeroshot setting trained on our synthetic Q-A pairs. The performance stagnates after a critical threshold of 10^6 samples is reached. Our experiments also suggest that randomly sampling a set of questions for each image per epoch leads to a $+4\%$ gain, as compared to training on the entire set.

D.7 Error Analysis

Our ZSL method is pretrained on longer phrases and hence tends to generate answers with more details, such as “red car" instead of “car". Although the SWA loss mitigates this to an extent, by creating a distribution over the shorter phrases, the bias is not completely removed. On automated evaluation, we observe that for 42% of questions the target answer is a sub-phrase of our predicted answer. Manual evaluation of 100 such samples shows that 87% of our detailed predictions are also plausible answers. This not only shows the relevance of learning from captions, but also quantifies the bias towards short “true" answers in human-annotated benchmarks, demonstrating the need for better evaluation metrics that do not penalize VQA systems for producing descriptive accurate answers.

In the fully supervised setting, we either finetune our pre-trained QA classifier with the SWA Loss, or train a separate feedforward layer for the task. The pre-trained QA classifier continues to predict longer phrases as answers, leading to a drop in accuracy. The feedforward layer (trained from scratch) performs better ($+6\%$), indicating our Encoder captures relevant features necessary to generalize to the benchmark answer-space. Note that we do not use object annotations during training, unlike existing methods.