Q: How to Specialize Large Vision-Language Models to Data-Scarce VQA Tasks? A: Self-Train on Unlabeled Images!

Zaid Khan[†*] Vijay Kumar BG[♣] Samuel Schulter[♣] Xiang Yu[♢*] Yun Fu[†] Manmohan Chandraker[♣♡]
[†]Northeastern University, [♣]NEC Labs America, [♢]Amazon, [♡]UC San Diego

Abstract

*Finetuning a large vision language model (VLM) on a target dataset after large scale pretraining is a dominant paradigm in visual question answering (VQA). Datasets for specialized tasks such as knowledge-based VQA or VQA in non natural-image domains are orders of magnitude smaller than those for general-purpose VQA. While collecting additional labels for specialized tasks or domains can be challenging, unlabeled images are often available. We introduce SelTDA (**Self-Taught Data Augmentation**), a strategy for finetuning large VLMs on small-scale VQA datasets. SelTDA uses the VLM and target dataset to build a teacher model that can generate question-answer pseudolabels directly conditioned on an image alone, allowing us to pseudolabel unlabeled images. SelTDA then finetunes the initial VLM on the original dataset augmented with freshly pseudolabeled images. We describe a series of experiments showing that our self-taught data augmentation increases robustness to adversarially searched questions, counterfactual examples and rephrasings, improves domain generalization, and results in greater retention of numerical reasoning skills. The proposed strategy requires no additional annotations or architectural modifications, and is compatible with any modern encoder-decoder multimodal transformer. Code available at* https://github.com/codezakh/SelTDA.

1. Introduction

Large, pretrained vision language foundation models [3, 20, 25, 26, 35, 49] are approaching human-level performance on visual question answering (VQA) [26, 50–52, 54, 62], as measured by the standard VQAv2 [13] benchmark. Yet on more complex VQA tasks [37, 43] there is a larger gap between humans and machines. One difficulty is the small scale of datasets for complex VQA tasks or those in domains beyond natural images. The first solution to deal with the data scarcity is to employ transfer learning from a larger VQA dataset (e.g. VQAv2) to the smaller, specialized VQA dataset. However weaknesses of VQA models such as lack of consistency [44], weakness to adversarially searched questions [27] and tendency to cheat by learning shortcuts [8] can be exacerbated when fine-tuning on small datasets.

*work done while at NEC Labs America

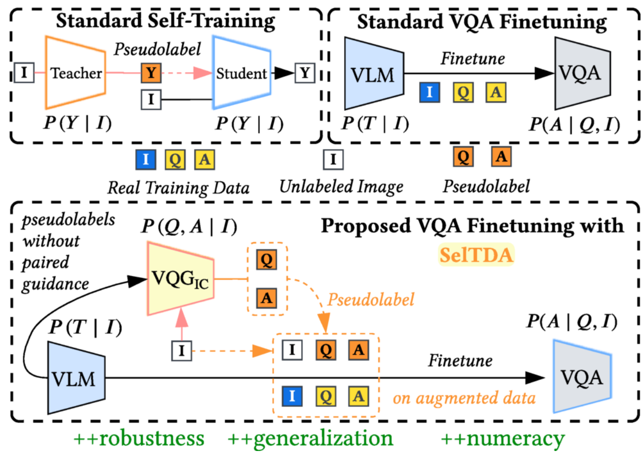

Figure 1. *SelTDA* expands the self-training paradigm to VQA. By self-generating supervision (orange line) for an image I without needing extra annotations, we can augment a target dataset with new images and their pseudo-questions and answers (Q, A).

Collecting annotations to expand a dataset for knowledge-intensive tasks or specialized domains is often prohibitively expensive. However, *unlabeled images* are cheap and often available. How can we exploit unlabeled images for specific visual question answering tasks? One possibility is to generate new question+answer pairs for the unlabeled images, and use them during training. However, existing methods for visual question *generation* require images with annotations — either ground truth captions [2, 4], or bounding boxes [21, 48]. Even if these annotations were to be acquired, they induce a limited set of possible questions; they are limited to objects and concepts included in the acquired annotation, which are in turn limited by the finite label space of pretrained object detectors and the information disparity between a caption and an image (an image usually contains much more content

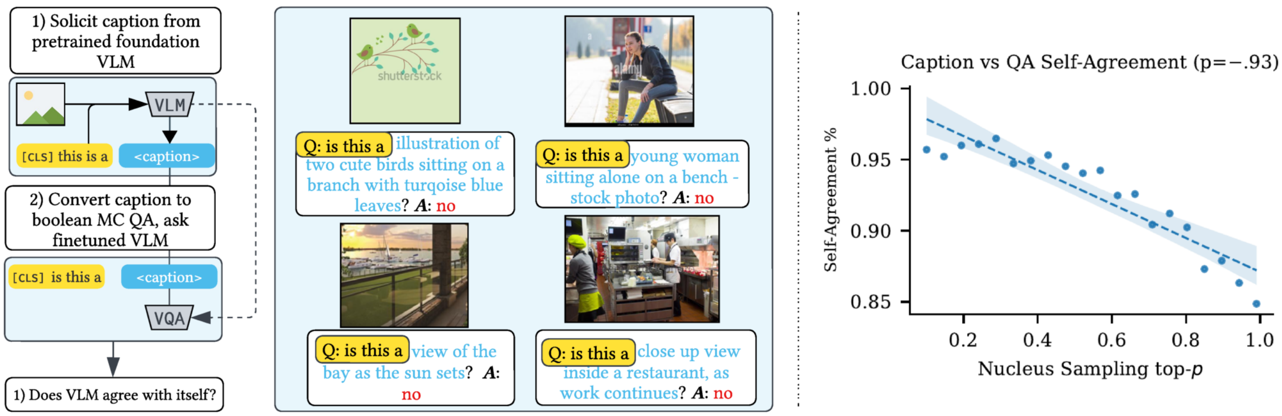

Figure 2. Motivating experiment. We sample increasingly diverse captions from BLIP [26], convert them to questions, and pose the questions to BLIP after finetuning on VQAv2. As caption diversity increases, self-agreement decreases (right panel). Despite the diversity, many captions remain correct (middle panel), suggesting that the VLM has knowledge that is not exhausted by task-specific finetuning.

than a short caption can describe).

Motivating Experiment: In Fig 2, we show that a large vision-language model (VLM) pretrained on web-scale data contains knowledge that can be drawn out with image-conditional text generation, but which the model cannot verify when posed as a visual question-answering task. We prompt the BLIP [26] VLM (pretrained on 129M image-text pairs) to caption 1000 images from the CC3M [45] dataset starting with the phrase "`this is a`". We convert each caption into a boolean question where the correct answer is "yes" by inserting the caption into the template `is this a <caption>?` Next, we ask a BLIP VLM finetuned on the VQAv2 dataset [13] to choose between "yes" and "no" for each caption turned into a question. Surprisingly, the VQA-finetuned BLIP answers "no" to *at least* 5% of the questions, increasing to 15% as the diversity of captions increases (adjusted by top-p parameter in nucleus sampling). This suggests the possibility that the VLM has knowledge it cannot exploit when answering questions, but is accessible when directly generating text conditioned on an image.

Approach: To exploit unlabeled images for VQA, we propose *SelTDA*, a three-stage framework for **Sel**f-**T**aught **D**ata **A**ugmentation (Fig 1 bottom panel). We adapt the paradigm of self-training used in object detection [29, 65] and image classification [41, 60] for VQA. In classification / detection, the task of labeling an image is identical to prediction, and the teacher and student optimize identically structured objectives. In VQA self-training, the student and teacher tasks are different. A teacher must pose and answer a question given an image, while the student provides an answers given a question and image. To handle this, we first cast the task of the teacher as a direct image-to-text generation task, and introduce a teacher model by updating the weights of the VLM to learn an *image-conditional* visual question generation model VQG_{IC}. Next, we use VQG_{IC} as a teacher to pseudolabel unlabeled images by sampling questions and answers from VQG_{IC} with stochastic decoding. Finally, we augment the original VQA dataset with the newly labeled image-question-answer pairs, and finetune the VLM for visual question answering on the augmented VQA dataset.

Benefits: *SelTDA* allows us generate synthetic training data by approximating the distribution $P(Q, A|I)$ of the target VQA task, where Q, A, I represents a question, answer, and image respectively. One benefit is that the synthetic data increases the number of training pairs available for finetuning, which effects an increase in raw performance. A second benefit is an increase in the diversity of questions and answers due to the introduction of new images and the stochastic nature of the text decoding, which results in increased robustness and domain generalization. A third benefit is the distillation of knowledge from pretraining and transfer learning into the synthetic training data, which can teach new skills (e.g. domain generalization) or prevent the forgetting of specific skills (e.g. numerical reasoning). Finally *SelTDA* is architecture-agnostic given a vision-language model capable of image-conditional text-generation. Our contributions can be summarized as follows:

1. We introduce *SelTDA*, a variant of the self-training paradigm that is designed for VQA and large generative pretrained VLMs.
2. We propose treating visual question generation as a direct image-to-text task by leveraging the autoregressive decoder of a large, pretrained VLM, enabling us to generate questions and answers from an unlabeled image with no auxillary annotations needed.
3. We show that a large VLM trained with the proposed *SelTDA* gains increased robustness, domain generalization, numerical reasoning, and performance when finetuning on small-scale VQA datasets.

2. Related Work

Augmentation for VQA The method of [53] augments images by using an MLP to classify possible answers in the image and using an LSTM to generate questions matching the answer. While this works with unlabeled images, it is not used for self-training, has a limited label space, and does not leverage large VLMs. KDDAug [6] augments existing question answer pairs by generating pseudoanswers and achieves increases in robustness. ConCat [19] similarly trains more robust models by augmenting the *existing* QA pairs in a dataset. In contrast to this line of work, we seek to exploit *unlabeled images* by generating *new* questions and answers, and using a large VLM to generate augmentation.

Few/Zero-shot Generalization Large VLMs have shown impressive generalization to unseen tasks after large-scale pretraining [1], echoing similar achievements in natural language processing [7, 55]. We explore zero-shot generalization to similar tasks in new domains. Domain *adaptation* in VQA has been explored, first by [5, 58] and most recently by [63]. These fall into the general line of *feature adaptation* methods for domain adaptation, as they align domain features. Our method is more similar to pseudolabeling based methods for domain adaptation [24, 31] with the difference being that our pseudolabels are natural language rather than distributions. Moreover, we do not focus on *adaptation*, but zero-shot generalization.

Visual Question Generation is a well-explored topic with a long history of prior work [23, 28, 38, 64]. In contrast to prior work, our VQG teacher model *does not* rely on or need paired ground truth annotations for an unlabeled image to generate questions. SimpleAug [21] and GuidedVQG [48] relies on annotations such as bounding boxes to generate new questions, and requires pretrained object detectors, which have a limited label space. WeaQ [2] requires captions to already be present, as does [4], which additionally uses a large language model (T5-XXL with 11B parameters) to generate questions. One similarity of our approach to [4] is that we both seek to use knowledge in a large model to generate questions, with the main differences being that we do not require ground-truth captions for unlabeled images, and we use a large vision-language model than a large language model. VQAPG [61] is similar to our approach in not requiring any ground-truth annotations, but focuses on creating a joint question-generation and question answering model that is consistent, rather than self-training a model with unlabeled data. The authors of [17] propose a VQG method that does not rely on ground-truth annotations, but their method is LSTM-based, rather than based on self-training with a large vision-language model.

Self-Training uses labeled data to train a teacher model. The teacher model provides labels for auxiliary unlabeled data. Finally, a student model is trained on the labeled data augmented with newly-labeled data. Previous work in self-training for computer vision focuses on image-classification [57, 59] or object detection [29, 41, 60, 65]. A significant difference between classical self-training and our setting is that in the more traditional settings, the teacher and student have the same task. In our setting, the task of the teacher (ask a question) is different than the task of the student (answer a question). More similar to us, [42] uses self-training for question-answering. However, the teacher model of [42] has a fundamentally different task, since it is a reading comprehension task, where the ground-truth answer is mentioned within the passage itself. In our task, the teacher model must generate the ground-truth answer from its own internal knowledge and by inspecting an image.

3. Method

Our goal is to pseudolabel an unlabeled image I with a generated question-answer pair (Q, A) using a teacher (initialized from the VLM), and then train a student model (the initial VLM) on the real VQA pairs augmented with the generated VQA pairs. To generate the pseudolabels, we first learn a visual question generation model on the real question-answer pairs and images as the teacher. We denote this model $\mathrm{VQG}_{\mathrm{IC}}$ to highlight the *image-conditional* nature of the model, because the model generates both a question and answer conditional on an image alone. This approach is end-to-end, requires *no ground truth annotations, bounding boxes, or handcrafted guidance*, and provides a generative model approximating $P(Q, A|I)$ that we can sample from. We then feed the teacher model unlabeled images and stochastically decode from the teacher model to generate pseudolabels, which we parse into question answer pairs. After the real samples in the dataset have been augmented with the self-generated samples, VQA training can proceed as normal. Our approach is compatible with any modern encoder-decoder multimodal architecture. This is because our approach relies entirely on direct image-to-text generation, which is possible in modern large vision language models since their autoregressive decoders are designed to produce text conditioned on an image.

3.1. The Teacher: Direct Image-Conditional VQG

Self-training requires a teacher model to produce pseudolabels that the student model then learns to mimic. In order to use unlabeled data for VQA, the teacher model must be able to pose a question and provide an answer given an unlabeled image, which is a different task from VQA. Given an image I, a question Q and answer A, the VQA student must approximate $P(A \mid Q, I)$, while the teacher model must approximate $P(Q, A \mid I)$. Previous approaches to visual question generation (VQG) cannot work with unlabeled data because they approximate $P(Q \mid I, A)$, that is, they generate a question conditional on the image and a potential answer. In contrast to these previous, answer-conditional VQG

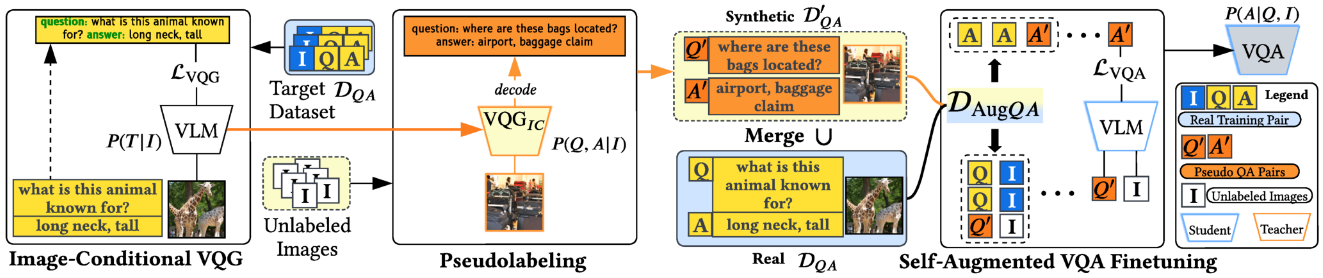

Figure 3. Overview of the proposed framework. We first create the teacher VQG_{IC} (§3.1), use VQG_{IC} to pseudolabel unlabeled images (§3.2), and finetune student on the original training pairs augmented with the pseudolabeled images. The pseudolabels are natural language.

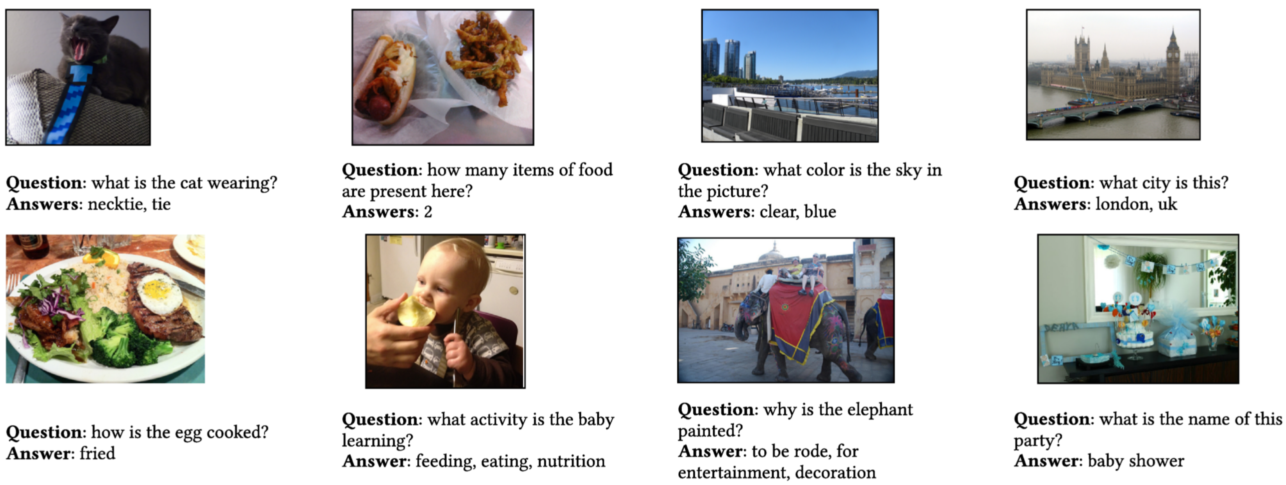

Figure 4. Example questions and answers generated by the teacher on unlabeled images. The questions include unusual pairings (cat wearing necktie) or require broad knowledge (identifying a baby shower or London landmarks) and inferences about scenes (the baby is learning).

approaches, we develop an *image-conditional* approach (VQG_{IC}) that we use as a teacher model. Our approach also contrasts with self-training in image classification or object detection, which benefit from having the teacher and student *both* approximating and predicting identically structured distributions $P(Y|I)$, where Y is often a distribution over a (finite) label space.

To create the VQG_{IC} teacher that approximates $P(Q,A|I)$, we treat the problem of learning such a model as a text-generation problem, and wish to train the autoregressive decoder of the vision-language model to approximate $P(T|I)$, where $T = (Q,A)$. Let $\mathcal{D}_{QA}$ be a question-answer dataset we wish to create a teacher from. For a sample $(Q,A,I) \in \mathcal{D}_{QA}$, we transform it into a target sequence of tokens $y_{1:N} = (y_1, y_2, \dots y_n)$ by entering (Q,A) into a structured template of the form "**Question**: <question>**? Answer**: <answer>." where <question> and <answer> are replaced by the content of Q and A respectively. Once $y_{1:N} = (y_1, y_2, \dots y_n)$ is obtained, we train the model by optimizing

$$\mathcal{L}_{\text{VQG}} = -\sum_{n=1}^{N} \log P_\theta\left(y_n \mid y_{<n}, x\right) \tag{1}$$

over all question-image-answer pairs in $\mathcal{D}_{QA}$, where x is the latent encoded features in the standard encoder-decoder architecture and θ represents the VLM parameters. The VQG_{IC} thus learns to maximize the conditional likelihood of a question-answer *pair* represented as a unified string, given an image. Recall that VQG_{IC} is initialized from the parameters of an autoregressive VLM. The VLM is a quality approximator of $P(T|I)$, having been exposed to a diverse number of images and paired text. The VQG_{IC} teacher can tap into this reservoir of knowledge, because a pseudo question-answer pair (Q', A') is generated jointly as a text T', allowing us to sample from $P(T|I)$.

3.2. Training the Student with Unlabeled Data

Once the VQG_{IC} teacher model has been obtained, self-training with unlabeled data can proceed. To produce a pseudolabel (Q', A') for an unlabeled image I_u, we first obtain $\mathbf{L}_{1:N} = \text{VQG}_{\text{IC}}(I_u)$, where $\mathbf{L}_{1:N}$ are the logits of the decoder. The logits $\mathbf{L}_{1:N}$ define a distribution $P(L_N \mid L_{1:N-1})$ over the tokens of the model's natural language vocabulary. We then apply nucleus sampling [15] to stochastically decode a text T' from $P(L_N \mid L_{1:N-1})$. The structured format of the generation template can then be easily parsed by a regular expression to recover a pseudo-question-answer pair (Q', A') from the decoded text T'. This pair $(Q', A') = T'$ is a sample from $P(T|I)$, and reflects textual knowledge about the content of an image known to the VLM.

We then proceed to pseudolabel the desired number of images and obtain any number of triplets of the form (Q', A', I_u), representing self-generated training data $\mathcal{D}'_{QA}$in the style of a target dataset $\mathcal{D}_{QA}$. We then augment the real dataset $\mathcal{D}_{QA}$ with the self-generated question-answer pairs on unlabeled images $\mathcal{D}'_{QA}$ to create a self-augmented training dataset $\mathcal{D}_{\text{Aug}QA} = \mathcal{D}'_{QA} \cup \mathcal{D}_{QA}$. The teacher model is no longer needed, and the student can be initialized from the checkpoint obtained after large-scale pretraining that the teacher model was initialized from. At this point, VQA training can proceed as normal. In our setting, we use the training procedure of BLIP [26] in which VQA is treated as an open-ended generation task, and the VQA objective can be expressed as the standard language modeling loss

$$\mathcal{L}_{\text{VQA}} = -\sum_{n=1}^{N} \log P_\theta (y_n \mid y_{<n}, x_n) \tag{2}$$

where x_n is the n-th element of the multimodal sequence embeddings $\mathbf{X}_{1:N}$ produced by $\text{VLM}(Q, I; \theta))$, Q, I are the question and image, $y_{1:N}$ is the sequence of answer tokens, and θ represents the VLM parameters, which we initialize from the *pretrained* weights rather than the teacher. Why can high quality pseudolabels (Q', A') be generated even when $\mathcal{D}_{QA}$ is small, and few pairs are available for adapting the teacher VQG_{IC}? Knowledge about the *content* of the image in a textual form $P(T|I)$ is already well-learned by the VLM from which we initialize VQG_{IC}. Thus, $\mathcal{D}_{QA}$ only needs sufficient pairs to teach VQG_{IC} how to construct annotations matching the style of $\mathcal{D}_{QA}$.

4. Experiments

Experimental Setup We implement our framework in PyTorch [39] and use the same hyperparameter settings for all experiments. Our settings are taken from [26]. We train each VQA model for 10 epochs, using the AdamW [33] optimizer with a weight decay of 0.05 and a linear LR decay to 0 from an initial LR 2e-5. Each VQG model is trained for 10 epochs with the same weight decay and an initial LR of 2e-5. For VQA, we use a global batch size of 64 on 4 GPUs, with a per device batch size of 16. For VQG, we use a global batch size of 128, with a per device batch size of 32. All models are initialized from pretrained BLIP [26] checkpoints. For VQA, we use an image size of 480×480 and an image size of 384×384 for VQG. For all datasets, we use the official training, validation, and test splits.

Baseline As a strong baseline model, we use the ViT-B/16 version of the BLIP [26] model pretrained on 129M image-text pairs. BLIP [26] has an autoregressive decoder and is trained for text-generation, making it easy to adapt to text-generation tasks. When decoding, we use nucleus sampling with a top-p of 0.92. Additional experiments and visualizations can be found in the supplemental material.

	Model	A-OKVQA Validation	A-OKVQA Test
(a)	ViLBERT [34]	49.1	41.5
(b)	LXMERT [46]	51.4	41.6
(c)	KRISP [36]	51.9	42.2
(d)	GPV-2 [18]	60.3	53.7
(e)	BLIP [26]	57.1	
(f)	$\text{BLIP}_{\text{VQAv2}}$ [26]	67.8	**59.5**
(g)	BLIP + *SelTDA*	62.1	54.5
	% gain w.r.t baseline	+5.0	
	% gain w.r.t best prior work	+1.8	+0.8
(h)	$\text{BLIP}_{\text{VQAv2}}$ + *SelTDA*	**68.9**	**59.5**
	% gain w.r.t baseline	+1.1	+0.0
	% gain w.r.t best prior work	+8.6	+5.8

Table 1. *SelTDA* improves performance on knowledge-based VQA, even on a strong baseline pretrained on 129M pairs.

	Model	ArtVQA Accuracy Overall	ArtVQA Accuracy Grounded
(a)	BAN [22]	22.4	-
(b)	BLIP [26]	21.36	81.71
(c)	VIKING [12]	55.5	78.74
(d)	$\text{VIKING}_{\text{VLM}}$	55.9	81.9
(e)	BLIP + *SelTDA*	21.68	**83.86**
	% gain w.r.t baseline	+0.32	+2.15
(f)	$\text{VIKING}_{\text{VLM}}$ + *SelTDA*	**56.86**	**83.86**
	% gain w.r.t baseline	+0.92	+1.96

Table 2. *SelTDA* improves VQA on fine art images [12] for VIKING and BLIP models. Grounded denotes visually grounded questions.

4.1. Self-Training: A-OKVQA & ArtVQA

We evaluate *SelTDA* in two domains: outside knowledge VQA on natural images with A-OKVQA [43] and outside knowledge VQA on fine-art images with AQUA [12]. We use the COCO 2017 unlabeled set [30] as a source of addi-

Question Type	Well-Posed Question	Answers Correct	Answerable	% of Total (95% CI)
External Knowledge	73%	62%	70%	29.6% - 50.00%
Visual Identification	94%	88%	94%	11.18% - 27.65 %
Visual Reasoning	83%	70%	80%	32.54% - 53.17%
Overall (95% CI)	71.16% - 87.96%	59.77% - 78.98%	68.83% - 86.22%	

Table 3. We manually inspect 100 questions and answers generated by the teacher model finetuned on A-OKVQA. We show the 95% confidence interval obtained by a proportion test. Annotator agreement on A-OKVQA is about 79.5% on the validation set.

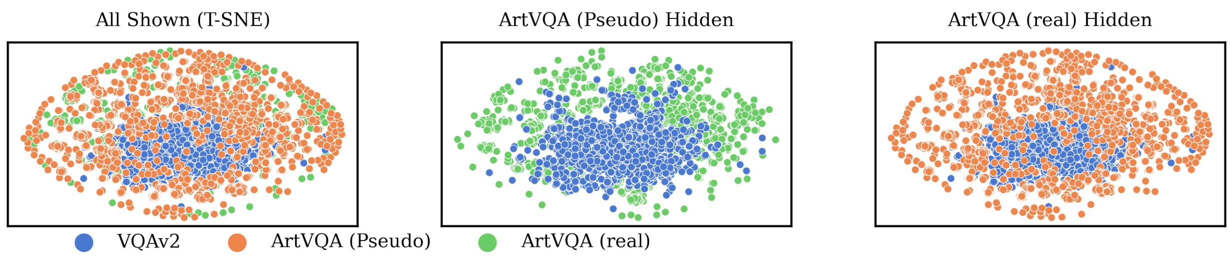

Figure 5. A T-SNE embedding shows that questions generated by a teacher finetuned on ArtVQA (orange) differ from real VQAv2 questions (blue) and are more similar to the real ArtVQA questions (green), yet more diverse, covering a larger area. We use SimCSE [10] to obtain a dense vector representation of each sentence. All the sets of questions are embedded together with T-SNE.

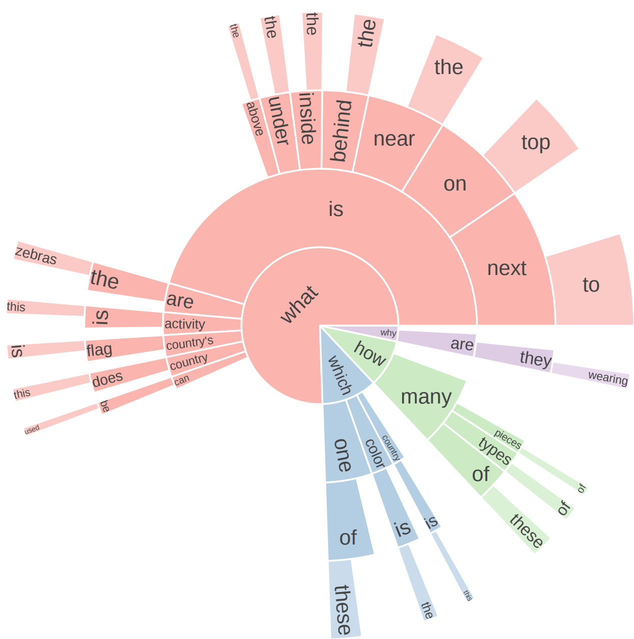

Figure 6. Sunburst chart of questions generated by a teacher model finetuned on A-OKVQA.

tional images for A-OKVQA, and SemArt [11] as a source of fine art images for ArtVQA. On A-OKVQA, we perform model selection over students trained with varying amounts of *SelTDA* with the training set, and on ArtVQA, we use the validation set. On A-OKVQA (Table 1), we show that self-taught data augmentation improves overall performance, especially in the setting where no extra data (VQAv2) is available. BLIP with *SelTDA* achieves SOTA performance on A-OKVQA without transfer learning (row g in Table 1), even relative to competitors using transfer learning. This performance improvement holds even when $447k$ *real* pairs from VQAv2 are used for transfer learning, suggesting that self-taught data augmentation offers real improvements over manual annotations. On fine art VQA (Table 2), we show that self-taught data augmentation achieves state-of-the-art and improves overall performance, with a large increase for visually grounded questions.

4.2. Ablations & Analysis of Pseudolabels

We manually evaluate 100 randomly sampled questions generated by the teacher model on A-OKVQA (Table 3). The generated questions and answers are noisier than the real questions and answers, but the levels of noise are not substantially below the human agreement on A-OKVQA. Questions which require visual reasoning or external knowledge are harder to generate correctly compared to those that require simpler visual identification (e.g. "what is this object?"). Next, we show using t-SNE [47] that the teacher model learns to copy the "style" of questions in a particular dataset (Fig 5). Synthetic questions generated by a teacher finetuned for a specific dataset (ArtVQA) are more similar to the style of the questions found in the target dataset compared to real questions from a different dataset (VQAv2), while being more diverse.

Images		Questions						
Labeled	Unlabeled	Real	Synthetic	Total	Multiplier	Accuracy	% Gain	Questions/Image
17,000	0	17,000	0	17,000	1x (baseline)	57.11		N/A
17,000	0	17,000	17,000	34,000	2x	57.85	+0.74	1 / 1
17,000	0	17,000	34,000	51,000	3x	**60.01**	**+2.90**	2 / 1
17,000	0	17,000	51,000	68,000	4x	59.73	+2.62	3 / 1
17,000	0	17,000	0	17k	1x (baseline)	57.11		N/A
17,000	8,500	17,000	17,000	34,000	2x	60.69	+3.57	2 / 1
17,000	17,000	17,000	34,000	51,000	3x	**62.09**	**+4.98**	2 / 1
17,000	25,500	17,000	51,000	68,000	4x	61.31	+4.20	2 / 1

Table 4. *SelTDA* can improve performance even without additional unlabeled images, by generating more QA pairs for already labeled images. However, using previously unlabeled and unseen images results in further improvements. A-OKVQA is used.

	# of Real + Synthetic QA Pairs			Robustness Test Sets				
	Real	Synthetic	Multiplier	AdVQA	VQA-CE	VQA-Rephrasings	Avg. % Increase	Robustness Total
(a)	17,000	0	$\times 1$	31.06	51.43	65.88	0	148.37
(b)	17,000	2,000	$\times 1.1$	37.09	52.96	67.94	+3.21	157.99
(c)	17,000	4,500	$\times 1.3$	36.99	53.15	**67.98**	+3.25	158.12
(d)	17,000	8,000	$\times 1.5$	37.34	**53.33**	67.57	**+3.29**	**158.24**
(e)	17,000	12,000	$\times 1.7$	**37.43**	52.62	67.35	+3.01	157.4
(f)	17,000	17,000	$\times 2$	36.95	52.05	66.95	+2.53	155.95
(g)	17,000	34,000	$\times 3$	36.89	51.00	65.64	+1.72	153.53
(h)	17,000	51,000	$\times 4$	36.06	50.25	64.78	+0.91	151.09
	Max % increase on each dataset			+6.03	+1.9	+2.1		+9.87

Table 5. *SelTDA* improves robustness of VQA models on AdVQA (adversarially searched questions), VQA-CE (multimodal shortcut learning) and VQA-Rephrasings test sets. The baseline (a) is trained on VQAv2 after pretraining, then finetuned on A-OKVQA.

We show that the performance gains of *SelTDA* are due to novel-question answer pairs (first half of Tab 4) that add information not present in the ground-truth QA pairs, not only due to the additional images. However, the student model benefits from *both* the novel-question answer pairs and unlabeled images (second half of Table 4).

Optimal Amount of Augmentation We explore how the amount of augmentation affects performance. The highest performance on the A-OKVQA validation and test sets is reached when the number of synthetic is double that of the real pairs (Table 4). When transfer learning from VQAv2, the ratio is different, and peak performance is reached when the number of synthetic pairs is 50% the number of real pairs (Table 5,4). Performance and robustness improvements (Table 5) saturate as increasing amounts of synthetic pairs are added, which may be the result of task-irrelevant information seeping into the dataset due to stochastic sampling.

4.3. Robustness

We investigate whether the self-taught data augmentation improves robustness of VQA models. We consider three known weaknesses. The first is adversarially searched questions, collected in the AdVQA [27] dataset through human-in-the-loop attacks against state-of-the-art VQA models. In Table 5, we show that models trained with self-taught data augmentation perform significantly better (20% relative improvement and 6% absolute improvement) on AdVQA. The second form of robustness we consider is resistance to multimodal shortcut learning, which the VQA-CE (Counterexamples) [8] test set measures. The test set is constructed so that models which have learned to answer questions using shortcuts based on correlations in the VQAv2 training set (ex: tennis racket detected + question about sport $\rightarrow$ always answer tennis) will display reduced performance on the VQA-CE test set. We construct our A-OKVQA models by transfer learning from the VQAv2 training set, so VQA-CE can be used to test multimodal shortcut learning in our models. In Table 5, we show that models trained with self-taught data augmentation are more resistant to shortcut learning (1.9% absolute improvement on VQA-CE) compared to the baseline model trained without self-taught data augmentation. Finally, we consider robustness to rephrasings. VQA models have been shown to be inconsistent when evaluated on rephrasings [44]. The VQA-Rephrasings test set consists of 3 human-provided rephrasings of the questions in the VQAv2 test set, intended to test the robustness of the model to rephrasings. On VQA-Rephrasings, self-taught data augmentation induces a 2.1% performance improvement relative to the baseline model, though both the baseline model and augmented models were initialized from from the same weights learned on the VQAv2 training set prior to finetuning on A-OKVQA.

4.4. Domain Generalization

We hypothesize that self-taught data augmentation may improve domain generalization, because the student model has been exposed to a greater diversity of questions and answers. To test this, we compare the generalization of the baseline model and models trained with self-taught data augmentation on unseen test sets from three different domains. Concretely, we treat the natural-image based A-OKVQA task as the source task, and evaluate on VQA datasets from three target domains: medical, fine art, and remote sensing. For medical VQA, we use the PathVQA [14] dataset containing question and answers on pathology images. For fine art, we used the previously described AQUA [12] dataset for visual question answering on art images. For remote sensing, we use the RSVQA dataset [32], containing question and answers on satellite images. We display the results in Table 6. Across all three domains, self-taught data augmentation improves domain generalization over the baseline model. The improvement is greatest on fine art images, as the fine art domain is closest to the natural image domain with respect to the images, questions, and answers.

4.5. Numerical Reasoning

Numerical reasoning is required to answer questions such as "how many sheep are looking at the camera". Naive transfer learning from VQAv2 to A-OKVQA results in catastrophic forgetting of numerical reasoning, and naive finetuning on A-OKVQA results in models with poor numerical reasoning. In Table 7, we show that *SelTDA* significantly aids numerical reasoning when finetuning on a small-scale VQA dataset such as A-OKVQA. We measure numerical reasoning using questions labeled as requiring numerical answers on VQAv2 and the VQA-Rephrasings datasets. When transfer learning from VQAv2 (first half of Table 7), self-taught data augmentation results in an absolute increase of 29.81% and 24.71% on numerical questions on VQAv2 and VQA-Rephrasings. When finetuning directly on A-OKVQA (2nd half of Table 7), self-taught data augmentation results in an absolute increase of 3.63% and 10.57%. These results suggest that self-taught data augmentation can prevent catastrophic forgetting of numerical reasoning when transfer learning, and improve numerical reasoning significantly,

	Target (0-shot)		
Model	ArtVQA	PathVQA	RSVQA
Baseline (BLIP)	31.65	25.09	37.78
BLIP + *SelTDA*	38.03	26.76	38.99
% gain w.r.t baseline	+6.38	+1.67	+1.1

Table 6. *SelTDA*improves domain generalization from natural images (A-OKVQA) to art QA, medical QA, and remote sensing QA.

	# Training Pairs		Numerical Reasoning	
Initialization	Real	Synth	VQAv2	VQA-Rephrasings
BLIP_{VQAv2}	17000	0	13.49	13.06
BLIP_{VQAv2}	17000	2000	38.73	33.74
BLIP_{VQAv2}	17000	4500	40.4	35.91
BLIP_{VQAv2}	17000	8000	42.9	36.5
BLIP_{VQAv2}	17000	12000	**43.3**	**37.77**
max % gain w.r.t baseline			+29.81	+24.71
BLIP	17000	0	1.42	1.29
BLIP	17000	17000	4.53	11.44
BLIP	17000	34000	**5.05**	11.77
BLIP	17000	51000	4.26	**11.86**
max % gain w.r.t baseline			+3.63	+10.57

Table 7. *SelTDA* improves numerical reasoning when finetuning on a small-scale dataset (A-OKVQA). BLIP_{VQAv2} indicates transfer learning from VQAv2, and BLIP indicates direct finetuning.

even when the dataset used to train the teacher model has few numerical reasoning questions. One reason for this is that the the word "how" is a high-probability word to start a question with, and is naturally followed by "many" (Fig 6) resulting in numerical questions being generated.

5. Conclusion & Future Work

We present *SelTDA*, a framework for self-improving large VLMs on small-scale visual question answering tasks with unlabeled data. The limitations of *SelTDA* suggest several opportunities for further work. First, the pseudo-QA pairs can be noisy. Combining *SelTDA*with methods for fact-checking based on external knowledge [40], logically consistent self-reasoning [16], or chain-of-thought prompting [56] to rationalize answers may result in higher quality pairs for self-training. Second, learning the teacher model may fail for specialized domains (e.g. medical), because the vocabulary is too specialized. Third, biases in the VLM or pretraining data may be amplified by self-training, and addressing these biases may reduce multimodal shortcut learning. Finally, self-training is yet to be explored with recently developed billion-parameter VLMs [9, 25].

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. *ArXiv*, abs/2204.14198, 2022. 3

[2] Pratyay Banerjee, Tejas Gokhale, Yezhou Yang, and Chitta Baral. Weaqa: Weak supervision via captions for visual question answering. In *FINDINGS*, 2021. 1, 3

[3] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, S. Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen A. Creel, Jared Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren E. Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas F. Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, O. Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir P. Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Benjamin Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, J. F. Nyarko, Giray Ogut, Laurel J. Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Robert Reich, Hongyu Ren, Frieda Rong, Yusuf H. Roohani, Camilo Ruiz, Jack Ryan, Christopher R'e, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishna Parasuram Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei A. Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models. *ArXiv*, abs/2108.07258, 2021. 1

[4] Soravit Changpinyo, Doron Kukliansky, Idan Szpektor, Xi Chen, Nan Ding, and Radu Soricut. All you may need for vqa are image captions. In *NAACL*, 2022. 1, 3

[5] Wei-Lun Chao, Hexiang Hu, and Fei Sha. Cross-dataset adaptation for visual question answering. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5716–5725, 2018. 3

[6] Long Chen, Yuhang Zheng, and Jun Xiao. Rethinking data augmentation for robust visual question answering. In *ECCV*, 2022. 3

[7] Hyung Won Chung, Le Hou, S. Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Wei Yu, Vincent Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc Le, and Jason Wei. Scaling instruction-finetuned language models. *ArXiv*, abs/2210.11416, 2022. 3

[8] Corentin Dancette, Rémi Cadène, Damien Teney, and Matthieu Cord. Beyond question-based biases: Assessing multimodal shortcut learning in visual question answering. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1554–1563, 2021. 1, 7

[9] Constantin Eichenberg, Sidney Black, Samuel Weinbach, Letitia Parcalabescu, and Anette Frank. MAGMA – multimodal augmentation of generative models through adapter-based finetuning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2416–2428, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics. 8

[10] Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. *ArXiv*, abs/2104.08821, 2021. 6

[11] Noa Garcia and George Vogiatzis. How to read paintings: Semantic art understanding with multi-modal retrieval. 2018. 6

[12] Noa Garcia, Chentao Ye, Zihua Liu, Qingtao Hu, Mayu Otani, Chenhui Chu, Yuta Nakashima, and Teruko Mitamura. A dataset and baselines for visual question answering on art. In *Proceedings of the European Conference in Computer Vision Workshops*, 2020. 5, 8

[13] 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 *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2

[14] Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020. 8

[15] Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *ArXiv*, abs/1904.09751, 2020. 5

[16] Jaehun Jung, Lianhui Qin, Sean Welleck, Faeze Brahman, Chandra Bhagavatula, Ronan Le Bras, and Yejin Choi. Maieutic prompting: Logically consistent reasoning with recursive explanations. *ArXiv*, abs/2205.11822, 2022. 8

[17] Kushal Kafle, Mohammed Yousefhussien, and Christopher Kanan. Data augmentation for visual question answering. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 198–202, Santiago de Compostela, Spain, Sept. 2017. Association for Computational Linguistics. 3

[18] Amita Kamath, Christopher Clark, Tanmay Gupta, Eric Kolve, Derek Hoiem, and Aniruddha Kembhavi. Webly supervised

concept expansion for general purpose vision models. *ArXiv*, abs/2202.02317, 2022. 5

[19] Yash Kant, A. Moudgil, Dhruv Batra, Devi Parikh, and Harsh Agrawal. Contrast and classify: Alternate training for robust vqa. *ArXiv*, abs/2010.06087, 2020. 3

[20] Zaid Khan, Vijay Kumar BG, Xiang Yu, Samuel Schulter, Manmohan Chandraker, and Yun Fu. Single-stream multi-level alignment for vision-language pretraining. In *ECCV*, 2022. 1

[21] Jihyung Kil, Cheng Zhang, Dong Xuan, and Wei-Lun Chao. Discovering the unknown knowns: Turning implicit knowledge in the dataset into explicit training examples for visual question answering. In *EMNLP*, 2021. 1, 3

[22] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. In *NeurIPS*, 2018. 5

[23] Ranjay Krishna, Michael S. Bernstein, and Li Fei-Fei. Information maximizing visual question generation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2008–2018, 2019. 3

[24] Ananya Kumar, Tengyu Ma, and Percy Liang. Understanding self-training for gradual domain adaptation. *ArXiv*, abs/2002.11361, 2020. 3

[25] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *ArXiv*, abs/2301.12597, 2023. 1, 8

[26] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 1, 2, 5

[27] Linjie Li, Jie Lei, Zhe Gan, and Jingjing Liu. Adversarial vqa: A new benchmark for evaluating the robustness of vqa models. In *International Conference on Computer Vision (ICCV)*, 2021. 1, 7

[28] Yikang Li, Nan Duan, Bolei Zhou, Xiao Chu, Wanli Ouyang, and Xiaogang Wang. Visual question generation as dual task of visual question answering. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6116–6124, 2018. 3

[29] Yandong Li, Di Huang, Danfeng Qin, Liqiang Wang, and Boqing Gong. Improving object detection with selective self-supervised self-training. *ArXiv*, abs/2007.09162, 2020. 2, 3

[30] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 5

[31] Hong Liu, Jianmin Wang, and Mingsheng Long. Cycle self-training for domain adaptation. In *NeurIPS*, 2021. 3

[32] Sylvain Lobry, Diego Marcos, Jesse James Murray, and Devis Tuia. Rsvqa: Visual question answering for remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 58:8555–8566, 2020. 8

[33] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. *ArXiv*, abs/1711.05101, 2017. 5

[34] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, 2019. 5

[35] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. *ArXiv*, abs/2206.08916, 2022. 1

[36] Kenneth Marino, Xinlei Chen, Devi Parikh, Abhinav Kumar Gupta, and Marcus Rohrbach. Krisp: Integrating implicit and symbolic knowledge for open-domain knowledge-based vqa. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14106–14116, 2021. 5

[37] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3190–3199, 2019. 1

[38] N. Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. Generating natural questions about an image. *ArXiv*, abs/1603.06059, 2016. 3

[39] 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 *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 5

[40] Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Dmytro Okhonko, Samuel Broscheit, Gautier Izacard, Patrick Lewis, Barlas Oguz, Edouard Grave, Wen-tau Yih, and Sebastian Riedel. The web is your oyster - knowledge-intensive NLP against a very large web corpus. *CoRR*, abs/2112.09924, 2021. 8

[41] Chuck Rosenberg, Martial Hebert, and Henry Schneiderman. Semi-supervised self-training of object detection models. In *2005 Seventh IEEE Workshops on Applications of Computer Vision (WACV/MOTION'05) - Volume 1*, volume 1, pages 29–36, 2005. 2, 3

[42] Mrinmaya Sachan and Eric Xing. Self-training for jointly learning to ask and answer questions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 629–640, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. 3

[43] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. *arXiv*, 2022. 1, 5

[44] Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. Cycle-consistency for robust visual question answering. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6642–6651, 2019. 1, 8

[45] 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 Association for*

Computational Linguistics (Volume 1: Long Papers), pages 2556–2565, Melbourne, Australia, July 2018. Association for Computational Linguistics. 2

[46] Hao Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *ArXiv*, abs/1908.07490, 2019. 5

[47] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. 6

[48] Nihir Vedd, Zixu Wang, Marek Rei, Yishu Miao, and Lucia Specia. Guiding visual question generation. *ArXiv*, abs/2110.08226, 2022. 1, 3

[49] 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. *CoRR*, abs/2202.03052, 2022. 1

[50] 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. *CoRR*, abs/2202.03052, 2022. 1

[51] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *ArXiv*, abs/2208.10442, 2022. 1

[52] 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. 1

[53] Zixu Wang, Yishu Miao, and Lucia Specia. Cross-modal generative augmentation for visual question answering. In *BMVC*, 2021. 3

[54] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *ArXiv*, abs/2108.10904, 2022. 1

[55] Jason Wei, Maarten Bosma, Vincent 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*, abs/2109.01652, 2022. 3

[56] 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*, abs/2201.11903, 2022. 8

[57] Qizhe Xie, Eduard H. Hovy, Minh-Thang Luong, and Quoc V. Le. Self-training with noisy student improves imagenet classification. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2020. 3

[58] Yiming Xu, Lin Chen, Zhongwei Cheng, Lixin Duan, and Jiebo Luo. Open-ended visual question answering by multi-modal domain adaptation. In *FINDINGS*, 2020. 3

[59] Ismet Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Kumar Mahajan. Billion-scale semi-supervised learning for image classification. *ArXiv*, abs/1905.00546, 2019. 3

[60] Qize Yang, Xihan Wei, Biao Wang, Xian-Sheng Hua, and Lei Zhang. Interactive self-training with mean teachers for semi-supervised object detection. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5937–5946, 2021. 2, 3

[61] Sen Yang, Qingyu Zhou, Dawei Feng, Yang Liu, Chao Li, Yunbo Cao, and Dongsheng Li. Diversity and consistency: Exploring visual question-answer pair generation. In *EMNLP*, 2021. 3

[62] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel C. F. Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, Ce Liu, Mengchen Liu, Zicheng Liu, Yumao Lu, Yu Shi, Lijuan Wang, Jianfeng Wang, Bin Xiao, Zhen Xiao, Jianwei Yang, Michael Zeng, Luowei Zhou, and Pengchuan Zhang. Florence: A new foundation model for computer vision. *ArXiv*, abs/2111.11432, 2021. 1

[63] Mingda Zhang, Tristan D. Maidment, Ahmad Diab, Adriana Kovashka, and Rebecca Hwa. Domain-robust vqa with diverse datasets and methods but no target labels. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7042–7052, 2021. 3

[64] Shijie Zhang, Lizhen Qu, Shaodi You, Zhenglu Yang, and Jiawan Zhang. Automatic generation of grounded visual questions. *ArXiv*, abs/1612.06530, 2017. 3

[65] Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin Dogus Cubuk, and Quoc Le. Rethinking pre-training and self-training. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 2, 3