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.

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 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).

*work done while at NEC Labs America
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 Self-Taught Data Augmentation (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 auxiliary 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] rely 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 rather 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 VQGIC 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 $P(A | Q, I)$, while the teacher model must approximate $P(Q, A | I)$. Previous approaches to visual question generation (VQG) cannot work with unlabeled data because they approximate $P(Q | I, A)$, that is, they generate a question conditional on the image and a potential answer. In contrast to these previous, answer-conditional VQG
approaches, we develop an image-conditional approach (VQGIC) 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 VQGIC 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 \(D_{QA}\) be a question-answer dataset we wish to create a teacher from. For a sample \((Q, A, I) \in D_{QA}\), we transform it into a target sequence of tokens \(y_{1:N} = (y_1, y_2, \ldots, y_N)\) 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 \(y_{1:N} = (y_1, y_2, \ldots, y_N)\) is obtained, we train the model by optimizing

\[
\mathcal{L}_{VQG} = - \sum_{n=1}^{N} \log P_{\theta}(y_n | y_{<n}, x)
\]

over all question-image-answer pairs in \(D_{QA}\), where \(x\) is the latent encoded features in the standard encoder-decoder architecture and \(\theta\) represents the VLM parameters. The VQGIC thus learns to maximize the conditional likelihood of a question-answer pair represented as a unified string, given an image. Recall that VQGIC 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 VQGIC 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 VQGIC 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 \(L_{1:N} = VQGI_C(I_u)\), where \(L_{1:N}\) are the logits of the decoder. The logits \(L_{1:N}\) define a distribution \(P(L_N | L_{1:N-1})\) over the tokens of the model’s natural language vocabulary. We then apply nucleus sampling [15] to stochastically decode a text \(T'\) from \(P(L_N | L_{1:N-1})\). 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 \(D_{QA}'\) in the style of a target dataset \(D_{QA}\). We then augment the real dataset \(D_{QA}\) with the self-generated question-answer pairs on unlabeled images \(D_{QA}'\) to create a self-augmented training dataset \(D_{AugQA} = D_{QA} \cup D_{QA}'\). 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

\[
\mathcal{L}_{VQA} = - \sum_{n=1}^{N} \log P_{\theta}(y_n | y_{<n}, x_n)
\]

where \(x_n\) is the \(n\)-th element of the multimodal sequence embeddings \(X_{1:N}\) produced by VLM(\(Q, I; \theta\)), \(Q, I\) are the question and image, \(y_{1:N}\) is the sequence of answer tokens, and \(\theta\) 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 \(D_{QA}\) is small, and few pairs are available for adapting the teacher VQG? 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 VQGIC. Thus, \(D_{QA}\) only needs sufficient pairs to teach VQGIC how to construct annotations matching the style of \(D_{QA}\).

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.

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-

<table>
<thead>
<tr>
<th>Model</th>
<th>Validation</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) BAN [22]</td>
<td>22.4</td>
<td>-</td>
</tr>
<tr>
<td>(b) BLIP [26]</td>
<td>21.36</td>
<td>81.71</td>
</tr>
<tr>
<td>(c) VIKING [12]</td>
<td>55.5</td>
<td>78.74</td>
</tr>
<tr>
<td>(d) VIKINGVLM</td>
<td>55.9</td>
<td>81.9</td>
</tr>
<tr>
<td>(e) BLIP + SelTDA</td>
<td>21.68</td>
<td>83.86</td>
</tr>
<tr>
<td>(f) VIKINGVLM + SelTDA</td>
<td>56.86</td>
<td>83.86</td>
</tr>
</tbody>
</table>

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

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

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.

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.
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). 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 → 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

---

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.

<table>
<thead>
<tr>
<th># of Real + Synthetic QA Pairs</th>
<th>Robustness Test Sets</th>
<th>Avg. % Increase</th>
<th>Robustness Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AdVQA</td>
<td>VQA-CE</td>
<td>VQA-Rephrasings</td>
</tr>
<tr>
<td>(a) 17,000 0 x1</td>
<td>31.06</td>
<td>51.43</td>
<td>65.88</td>
</tr>
<tr>
<td>(b) 17,000 2,000 x1.1</td>
<td>37.09</td>
<td>52.96</td>
<td>67.94</td>
</tr>
<tr>
<td>(c) 17,000 4,500 x1.3</td>
<td>36.99</td>
<td>53.15</td>
<td>67.98</td>
</tr>
<tr>
<td>(d) 17,000 8,000 x1.5</td>
<td>37.34</td>
<td>53.33</td>
<td>67.57</td>
</tr>
<tr>
<td>(e) 17,000 12,000 x1.7</td>
<td>37.43</td>
<td>52.62</td>
<td>67.35</td>
</tr>
<tr>
<td>(f) 17,000 17,000 x2</td>
<td>36.95</td>
<td>52.05</td>
<td>66.95</td>
</tr>
<tr>
<td>(g) 17,000 34,000 x3</td>
<td>36.89</td>
<td>51.00</td>
<td>65.64</td>
</tr>
<tr>
<td>(h) 17,000 51,000 x4</td>
<td>36.06</td>
<td>50.25</td>
<td>64.78</td>
</tr>
</tbody>
</table>

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.
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, even when the dataset used to train the teacher model has few numerical reasoning questions. One reason for this is that 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


[18] Amita Kamath, Christopher Clark, Tammay Gupta, Eric Kolve, Derek Hoiem, and Aniruddha Kembhavi. Webly supervised
concept expansion for general purpose vision models. ArXiv, abs/2202.02317, 2022. 5


[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


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


[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


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.


