computer sciences and mathematics forum

MDPI

Proceeding Paper

Semi-Supervised Implicit Augmentation for Data-Scarce VQA †

Bhargav Dodla [1,2,‡], Kartik Hegde [1,*,‡] and A. N. Rajagopalan [1]

[1] Indian Institute of Technology, Madras 600036, India; dodla.bhargav@gmail.com (B.D.); raju@ee.iitm.ac.in (A.N.R)
[2] Microsoft India (R&D) Private Limited, Bengaluru 560103, India
* Correspondence: kartikhegde0611@gmail.com
† Presented at the 2nd AAAI Workshop on Artificial Intelligence with Biased or Scarce Data (AIBSD), Vancouver, BC, Canada, 26 February 2024.
‡ These authors contributed equally to this work and share first authorship.

Abstract: Vision-language models (VLMs) have demonstrated increasing potency in solving complex vision-language tasks in the recent past. Visual question answering (VQA) is one of the primary downstream tasks for assessing the capability of VLMs, as it helps in gauging the multimodal understanding of a VLM in answering open-ended questions. The vast contextual information learned during the pretraining stage in VLMs can be utilised effectively to finetune the VQA model for specific datasets. In particular, special types of VQA datasets, such as OK-VQA, A-OKVQA (outside knowledge-based), and ArtVQA (domain-specific), have a relatively smaller number of images and corresponding question-answer annotations in the training set. Such datasets can be categorised as data-scarce. This hinders the effective learning of VLMs due to the low information availability. We introduce SemIAug (**Sem**i-Supervised **I**mplicit **Aug**mentation), a model and dataset agnostic strategy specially designed to address the challenges faced by limited data availability in the domain-specific VQA datasets. SemIAug uses the annotated image-question data present within the chosen dataset and augments it with meaningful new image-question associations. We show that SemIAug improves the VQA performance on data-scarce datasets without the need for additional data or labels.

Keywords: visual question answering; vision-language models; semi-supervised augmentation

Citation: Dodla, B.; Hegde, K.; Rajagopalan, A.N. Semi-Supervised Implicit Augmentation for Data-Scarce VQA. *Comput. Sci. Math. Forum* **2024**, *9*, 3. https://doi.org/10.3390/cmsf2024009003

Academic Editors: Kuan-Chuan Peng, Abhishek Aich and Ziyan Wu

Published: 7 February 2024

1. Introduction

Visual question answering (VQA) is the task of answering a given question, with the question asked conditioned on the associated image. Even though the task of answering a question for a given image is fairly simple for humans, it is very difficult for a model to provide an accurate answer to the question. In the process, the model encounters sub-tasks like understanding the unimodal context of the given image and question, cross-modal understanding, and predicting the most suitable answer. The two major factors that have the potential to enhance the effectiveness of VQA are the model architecture and the data. VQA is a data-intensive task that requires substantial amounts of labelled training data. A good dataset must span an exhaustive set of scenarios in both the vision and language domains so that the model has enough examples to learn the correct associations between the image and text. Standard benchmark datasets like [1,2] are used for general-purpose VQA tasks. However, in the recent past, new VQA datasets have been proposed to challenge VQA models in various ways. They include out-of-domain [3], additional knowledge-based [4,5], and application-specific [6,7] datasets. The quantity of annotated data in the context of specialised or challenging VQA, and VQA on application-specific datasets is far lesser, which can affect the performance of the model on the VQA task.

Dataset annotation is time-consuming, expensive, and can occasionally be susceptible to errors. Data augmentation is an alternate strategy to enhance the dataset [8] by applying

transformations to the existing dataset. The main purpose of data augmentation is to introduce variations in the input data that the model might encounter in real-world scenarios for generalisation. This process makes the model more robust and reduces the chances of overfitting small-scale datasets. For the purpose of the VQA task, augmentations in the vision modality, such as geometric and photometric transformations, are not ideal since they distort the information in the scene. Text-based augmentation techniques like [9,10] are generation-based, which suffer from quality and diversity issues. The recent rapid improvements and adaptation of language models have given a boost to vision-language research. There has been tremendous progress in multimodal VLMs in recent years. Some of the popularly used VLMs are Vilbert [11], VisualBERT [12], CLIP [13], ALBEF [14], BLIP [15], BLIP-2 [16], BridgeTower [17], ALPRO [18], etc. Current VLMs demonstrate exceptional zero-shot performance with the help of large-scale pretraining [19,20], improved pretraining strategies, and better architectures. Using task-specific finetuning, the knowledge in the pretrained VLM is used for downstream tasks, such as image captioning, VQA, multimodal feature extraction, etc.

We propose a VLM-based image-question augmentation since VLMs are adept at understanding natural language descriptions for the visual data. In this work, we propose an augmentation strategy and a method to improve the VQA task on special types of datasets like [4,5,7], which are data-scarce, i.e., they have a fewer amount of annotated question-answer pairs per image. This augmentation approach uses VLMs by harnessing the implicit information present in the dataset in the form of unmatched but relevant image-question pairs.

After pairing a relevant question to an image, the answer has to be obtained to complete the VQA triplet. We adopt pseudo-labelling [21], a semi-supervised learning (SSL) strategy to obtain the answer for the augmented VQA triplets. Pseudo-labelling is a technique used in SSL to utilise the training on the labelled data to generate labels for the unlabelled data. These new labels are termed 'pseudo-labels' as they are labelled by the model. Multiple recent works in semi-supervised learning (SSL) related to the vision domain, such as [22–24], consider pseudo-labelling for the improvement of the performance of the model. We set up our augmentation technique as a semi-supervised problem that uses pseudo-labels (i.e., answers) from a downstream VQA model, finetuned on the original VQA dataset to help in the augmentation. We show that the relevance of the augmented questions for that image using the SemIAug strategy is almost on par with the originally annotated questions.

The major contributions of our work are as follows:

- We introduce SemIAug, a non-generative model, as well as a dataset-agnostic data augmentation approach that performs augmentation by only using the images and questions from within a dataset;
- We propose a simple and effective technique that employs VLMs for matching new images and questions and reuses the same VLM for answering. This helps improve the performance of VQA tasks on datasets with relatively fewer question-answer pairs, i.e., data-scarce datasets;
- We propose a computationally efficient image-question matching strategy by extracting the image and question features using frozen VLMs.

2. Related Works

Most of the augmentation strategies in the text modality for VQA are generation-based, where image-grounded questions are generated. Such augmentations can broadly be classified into the following two categories: template-based and free-form generation. Ref. [10] benchmarks compositional spatio-temporal reasoning where questions are generated based on preset templates. The templates are created from various scenarios and filled with objects and their properties, which are detected from the image using scene graphs (a hierarchical representation of a scene that can express the objects, attributes, and

relationships in the scene) to create questions for various scenarios. The answer is also obtained from the scene graph.

Ref. [9] proposes the use of automated speech recognition models to transcribe videos with narrations. Candidate answers are then chosen from these transcriptions, and free-form questions are generated conditioned on these answers. The problem with generation-based augmentation is that the quality of the generated data is not as good as human annotations, and the generated data may not accurately represent the underlying data distribution. Additionally, generative models are computationally expensive to train and require large amounts of data to produce good-quality results.

Some works that are not generation based are SimpleAug [25] and KDDAug [26]. SimpleAug works on the principle that many of the "unknowns" are indeed "known" within the dataset. It uses object detectors to recognise objects in the image. It uses a strict matching criterion to match new questions to images based on the detected objects and their attributes. KDDAug uses less stringent conditions than SimpleAug and further explores the approach proposed by SimpleAug. After obtaining all reasonable image-question pairs, KDDAug leverages multi-teacher knowledge distillation (KD) to generate "soft" labels. This approach was proposed not only to avoid human annotations but also to be more robust to both in-domain and out-of-domain settings. These works utilise off-the-shelf object detectors and bounding boxes for detecting objects or counting instances of an object, etc., to match questions and obtain the answers. This requires the use of external modules, which are limited by their performance and are also at risk of propagating errors, which lead to performance bottlenecks.

The work of Zaid Khan et al. [27] is the most closely related to our work in terms of using large VLMs for data augmentation in visual question answering. It proposes SelTDA, a framework for self-improving large VLMs on small-scale VQA tasks with unlabelled data. SelTDA uses VLMs and the target dataset to build a teacher model that can generate question-answer pseudo-labels as captions directly conditioned on the image alone. This work is generation-based, which creates new data from existing data, which can introduce challenges such as quality and diversity issues. The work is centred around using additional unlabelled images for performance improvement, which requires images from outside the dataset of interest.

What distinguishes our proposed approach, SemIAug, as unique and elegant, first, is its utilisation of the versatile nature of VLMs, both to extract representations for the images and questions (used as a uni-modal extractor). Second is its flexibility of reusing the same or different VQA model by finetuning on a VQA dataset for use as an answering model. Unlike generation-based augmentation techniques, our work augments images with human-quality questions since it intelligently uses the questions within the dataset, which are human-annotated. It is also self-contained as it does not need external modules like object detectors, scene graphs, or the curation of additional image/question data.

3. Methodology

To achieve the goal of expanding a dataset with human annotation level quality and diversity, we harness the vast amount of implicit information present within the dataset, which forms the fundamental premise of SemIAug. We find this information as potential pairs of images and questions $\{I, Q'\}$, which can be matched from images depicting similar scenes or objects. As the first step of SemIAug, we systematically match such potential positive pairs and expand the dataset to a more densely matched version of itself. After matching the new pairs, we obtain the answer to these pairs using an answering model to complete the augmented triplet $\{I, Q', A'\}$. Figure 1 shows the complete setup of the augmentation strategy. It highlights the reusability of VLMs for feature extraction and the finetuned VQA model as the answering model for the newly matched pairs. We next describe, in detail, image-question matching and answering tasks as part of SemIAug.

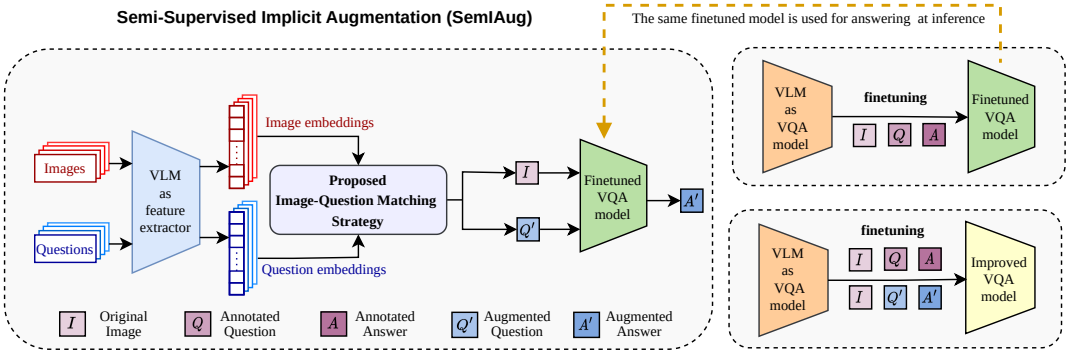

Figure 1. The block on the left depicts how the augmented $\{I, Q', A'\}$ triplets are obtained. The block at the top right depicts the finetuning of VLMs using the original data, and the block at the bottom right shows finetuning the VLMs with augmented data.

3.1. Image-Question Matching

We pose the image-question $\{I, Q'\}$ matching exercise as a question retrieval problem, i.e., use image and question embeddings and find the most relevant questions for the image using cosine-similarity. We use state-of-the-art VLMs to obtain these embeddings. Next, we explain the different building blocks of the image-question matching task.

3.1.1. Noun-Based Filtering

For a typical VQA dataset, the number of images and questions will be high. For example, if there are p images and q questions per image, then there are a total of $p \cdot q$ questions. In such a scenario, similarity-based retrieval between each image and all of the $p \cdot q$ questions will be very inefficient. We devise a simple yet effective filtering technique to reduce the potential set of questions Q to a more relevant and filtered set Q_f. From the set of questions annotated for a particular image, we extract all the nouns using the Python module NLTK [28], and use these as filters to restrict the set of possible augmentations to only those questions that contain these nouns. This helps in reducing the matching complexity from $O(MN)$ to $O(MK)$ where $K << N$.

3.1.2. Segregating Rephrased and Diverse Questions

In the filtered set of questions Q_f, there is the concern of rephrased questions for example, *'What is the man with the black hat doing?'* and *'What is the person wearing the black hat doing?'*. Though these are valid questions, they do not add any new information. Of course, they can be utilised to test for robustness since the answers to both questions must be the same. Note that our aim is to remove the rephrased versions of the originally annotated questions from Q_f. The embeddings obtained using the VLM feature extractor for the rephrased questions are expected to be similar. For separating the rephrased questions, we find the similarity between the feature vectors of annotated questions and the set of filtered questions Q_f (instead of every filtered question with one another). This reduces the complexity from $O(K^2)$ to $O(KP)$ where $P << K$. Questions with high cosine similarity with the original annotation are filtered out by setting an appropriate threshold (t). Thus, the questions that are left for matching are diverse questions Q_d that are best suited for augmentation.

3.1.3. Handling Multiplicity

Even after obtaining the diverse set of questions Q_d from the set of filtered questions Q_f, the number of potential questions to be augmented will typically still be large. For practical purposes, we require limiting the number of questions further. We can truncate the number of questions in two ways. One method is to apply a hard truncation common to all images such that the number of annotations n per image is $2\times$ or $3\times$ of the original count to obtain the final set of augmented questions Q_n per image. For example, if we are working

with *OK-VQA* or A-OKVQA datasets have been annotated with around one question per image in the training set, augmenting them with two or three questions will suffice. Another method would be a soft truncation to allocate the required question multiplicity image-wise. This approach dynamically assigns the count of augmentations per image in contrast to the previous setting, where we mandate a fixed number of augmentations to all images. This can be achieved by normalising the similarity score corresponding to all questions for an image with the highest similarity score. Then, a question relevance threshold can be applied, which will set the truncation count for the number of questions to be annotated for that image. This relevance threshold will set the augmentation multiplicity per image. For computational purposes, we set a minimum and maximum limit on how many questions can be dynamically allocated for an image.

3.2. Answering Model

After applying noun-based filtering, separating diverse questions, and truncating the number of augmented questions per image as part of image question matching, we end up with a combination of original and augmented image-question pairs. We have the human annotated answer $\{A\}$ for the original pairs $\{I, Q\}$ in the target dataset. To answer the newly matched question $\{Q'\}$, for the image $\{I\}$, we use the VLM trained on the same dataset as the VQA model. This is how our setup is analogous to a semi-supervised problem with the new dataset containing both the original (part of the dataset with labels) and new image-question pairs (part of the dataset without the answer labels). The answering model is trained on the original dataset, and using the trained model, we obtain the answers $\{A'\}$ (pseudo-labels) for the new image-question pairs. Then the original triplets $\{I, Q, A\}$ together with augmented triplets $\{I, Q', A'\}$ are used to form the new dataset for training. Finally, the VQA model is re-trained with the augmented dataset with the new triplets, which results in improved VQA accuracy.

4. Experimental Results

In this work, we utilised VLMs for both matching new questions and for answering. We used BLIP [15] and BLIP-2 [16], which are state-of-the-art VLMs, to obtain the unimodal embeddings of images and questions. We used frozen BLIP-2/BLIP image-text retrieval model checkpoint [15] pretrained on large amounts of data. We experimented with datasets OK-VQA and A-OKVQA as they are data-scarce (also requiring external knowledge) and, hence, good candidates to evaluate the efficacy of our proposed model. We re-used BLIP as the answering model, and all the other hyperparameters were kept the same as mentioned in [15]. We used 4 NVIDIA 3090 GPUs, each with 24GB RAM, to conduct our experiments.

First, we grouped all the training images and all the unique questions in the dataset and used BLIP-2/BLIP to extract the features. The feature extractor produces a 256-dimensional feature vector for each individual input image or question. For noun-based filtering, we adopted NLTK (as mentioned in Section 3.1.1) since it was faster. We then calculated the question-to-question cosine similarity using the question embeddings to analyse the effect of having rephrased questions versus a diverse set of augmented questions. We then set the threshold t such that all the questions that have a similarity score above t are considered to be rephrased versions of the original annotated questions. The rest constitute a diverse set of questions. We set t to be 0.8 based on our observations (please refer to the ablation Section 4.3). We calculated the image-question similarity and chose the *top-k* questions where k is a parameter that specifies the number of questions per image in the augmented dataset, which yields the augmented image-question pairs.

We used a semi-supervised setup for answering the augmented image-question pairs. As a strong baseline VQA model, we used the ViT-B/16 version of BLIP [15] pretrained on 129M image-text pairs and finetuned on VQA datasets like VQAv2 [2], Visual genome [29], provided by [15].

4.1. Quantitative Analysis

We finetuned the BLIP [15] baseline VQA model on the dataset of interest (OK-VQA or A-OKVQA) and re-used the finetuned model for answering the augmented image-question pair (obtained from BLIP-2 or BLIP). We finetuned the baseline VQA model for 10 epochs with an initial learning rate 2×10^{-6}. After obtaining the augmented triplets, we finetuned the same baseline VQA model but on the augmented dataset. We evaluated OK-VQA on its test set and, similarly, A-OKVQA on its designated val set. We observed that SemIAug improves VQA performance on both OK-VQA and A-OKVQA datasets over the baseline. We compared the accuracies of the BLIP VQA [15] baseline model with SemIAug with multiple fixed augmentation multiplicity, as presented in Table 1, and dynamic multiplicity, as presented in Table 2. We can observe that for all experiments with various multiplicities and both dynamic and fixed augmentation variants, SemIAug shows considerable performance gain over the baseline when using both BLIP-2 and BLIP as retrieval models. In Tables 1 and 2, we can observe close to or more than 1% improvement in the VQA accuracy on both the OK-VQA and A-OKVQA datasets using SemIAug, when using both BLIP-2 and BLIP. One critical observation across the fixed number and dynamic number of augmentations for comparable multiplicity when using BLIP retrieval, like $\times 2$ for OK-VQA in Table 1 and $\times 1.83$ for OK-VQA in Table 2, is that the dynamic augmentation has better performance, even with a slightly lower multiplicity, because the model chooses new questions when it passes the relevance threshold of 0.95. Hence, we can consider that it allows only relevant questions. A similar example can be found for A-OKVQA, where the accuracy of the model with a multiplicity $\times 3$ in Table 1 is lower than the multiplicity $\times 2.48$ in Table 2. With the higher relevance score, there is a higher chance of obtaining good image-question matches, but the number of newly matched image-question pairs will be reduced. Similar observations can be found when using BLIP-2 as the retrieval model. For example, for the OK-VQA dataset, a dynamic multiplicity of $\times 1.64$ is better than a multiplicity of $\times 2$, and for the A-OKVQA dataset, a dynamic multiplicity of $\times 2.18$ is better than a multiplicity of $\times 3$.

Table 3 provides a comparison of accuracy with various multimodal models on OK-VQA dataset. Using SemIAug, we can observe an improvement of 1.18% in accuracy when using BLIP as the retrieval model and an improvement of 0.96% in accuracy when using BLIP-2 as the retrieval model (both using finetuned BLIP for answering). We can also use the finetuned VQA checkpoint provided along with the VLMs for answering whenever available instead of finetuning the same. This would slightly deviate from the idea of SSL, as the provided checkpoint may be trained on more than just the target dataset.

Table 1. We compare the performance of SemIAug for various multiplicities with a fixed number of augmentations per image using BLIP 2 and BLIP as retrieval models and BLIP as the common VQA model.

Dataset	Model	Ques Count	Multiplicity	BLIP Retrieval		BLIP-2 Retrieval	
				Accuracy	% Gain	Accuracy	% Gain
OK-VQA (evaluated on test set)	Finetuned Baseline [15]	9000	$\times 1$	55.29	-	55.29	-
	SemIAug (Ours)	18,000	$\times 2$	55.74	0.45	55.78	0.49
		27,000	$\times 3$	56.27	0.98	55.83	0.54
		36,000	$\times 4$	**56.47**	**1.18**	**56.25**	**0.96**
A-OKVQA (evaluated on val set)	Finetuned baseline [15]	17,000	$\times 1$	54.35	-	54.35	-
	SemIAug (Ours)	34,000	$\times 2$	54.96	0.61	**55.77**	**1.42**
		51,000	$\times 3$	**55.48**	**1.13**	54.49	0.14
		68,000	$\times 4$	55.31	0.96	55.02	0.67

Table 4 provides a comparison of accuracy with various multimodal models on the A-OKVQA dataset. The assessment of A-OKVQA can be performed in two ways, either on direct answer (DA) or on multiple choice (MC) type, and here we chose DA. SemIAug with BLIP as retrieval improves the accuracy on the A-OKVQA dataset by 1.37% over the baseline implementation, whereas it shows an improvement of 1.42% when using BLIP-2 for retrieval. We have also experimented with sampled versions of the VQAv2 [2] dataset to simulate data-scarce settings, and this can be found in Appendix A.

Table 2. We compare the performance of SemIAug for various multiplicities with a dynamic number of augmentations per image using BLIP 2 and BLIP as retrieval models and BLIP as the common VQA model.

Dataset	Relevance Threshold	Lower Limit	Upper Limit	BLIP Retrieval			BLIP-2 Retrieval		
				Multiplicity	Accuracy	% Gain	Multiplicity	Accuracy	% Gain
OK-VQA (evaluated on test set)	-	-	-	×1	55.29	-	×1	55.29	-
	0.95	1	3	×1.83	**56.14**	**0.85**	×1.55	55.64	0.35
	0.9	1	3	×2.39	56.06	0.77	×2.02	55.72	0.43
	0.95	1	5	×2.10	55.60	0.31	×1.64	55.83	0.54
	0.9	1	5	×3.27	56.09	0.8	×2.45	**56.01**	**0.72**
A-OKVQA (evaluated on val set)	-	-	-	×1	54.35	-	×1	54.35	-
	0.95	1	3	×1.95	55.05	0.7	×1.67	55.08	0.73
	0.9	1	3	×2.48	**55.72**	**1.37**	×2.18	**55.37**	**1.02**
	0.95	1	5	×2.32	55.57	1.22	×1.82	55.02	0.67
	0.9	1	5	×3.56	55.66	1.31	×2.79	55.22	0.87

Table 3. Comparison of the accuracy of large multimodal models on OK-VQA. * indicates accuracy obtained by our implementation of the BLIP VQA baseline.

Model	Accuracy (Test)
(a) KAT (single) [30]	53.10
(b) REVIVE (single) [31]	56.60
(c) Unified-IO (2.8B) [32]	54.00
(d) ALBEF [14]	54.70
(e) BLIP [15]	55.40
(f) BLIP [15] *	55.29
(g) $BLIP_{retrieval}$ + $BLIP_{VQA}$ (SemIAug)	56.47
% gain w.r.t our baseline implementation (f)	+1.18
(h) $BLIP\text{-}2_{retrieval}$ + $BLIP_{VQA}$ (SemIAug)	56.25
% gain w.r.t our baseline implementation (f)	+0.96

Table 4. Comparison of the accuracy of large multimodal models on A-OKVQA. * indicates accuracy obtained by our implementation of the BLIP VQA baseline.

Model	Accuracy (Val)
(a) ViLBERT [11]	30.60
(b) LXMERT [33]	30.70
(c) GPV-2 [34]	48.60
(d) ALBEF [14]	54.50
(e) BLIP [15]	56.20
(f) BLIP [15] *	54.35
(g) $BLIP_{retrieval}$ + $BLIP_{VQA}$ (SemIAug)	55.72
% gain w.r.t our baseline implementation (f)	+1.37
(h) $BLIP\text{-}2_{retrieval}$ + $BLIP_{VQA}$ (SemIAug)	55.77
% gain w.r.t our baseline implementation (f)	+1.42

4.2. Qualitative Analysis

We conducted a user study to validate our claim that our image-question matching strategy provides new, good-quality, and relevant augmentations by utilising the existing information within the dataset. We designed a survey for assessing the relevance of newly matched image-question pairs, which were obtained by using BLIP image-text matching, against the originally annotated ones both on OK-VQA and A-OKVQA datasets. A total of 20 users participated in the study. We explained the nature of OK-VQA and A-OKVQA datasets to the user and asked them to rate the relevance of the questions for each image. For every participant, we sampled 20 images from each dataset and randomly picked a question corresponding to the image from the augmented dataset. We ensured that each user gets random image-question pairs and the information, whether the question was originally annotated or newly matched, is hidden. For the survey, each user had to rate the relevance of the question to the images on a scale of 1–5, with 5 being the best match and 1 being the poor match. Later, we found the weighted average relevance score by accounting for the number of originally annotated questions and newly matched questions that appeared for each user in the sampled data. Then, the average image-question relevance score is calculated over all the users for each OK-VQA and A-OKVQA dataset.

The results of the user study are shown in Figure 2. For the OK-VQA dataset, the average user relevance ratings for the originally annotated questions was 4.19, with a standard deviation of 0.42 and an average rating of 3.72 with a standard deviation of 0.47 for the newly matched image-question pairs. Similarly, for the A-OKVQA dataset, the average ratings for the original questions was 4.06, with a standard deviation of 0.27 and 3.61, with a standard deviation of 0.41 for the newly matched image-question pairs. We performed an one-sided analysis of variance (ANOVA) on the results obtained from the user study conducted on the OK-VQA and A-OKVQA datasets. The ANOVA yielded an F-statistic of 8.18 with a p-value of 0.0079 on OK-VQA user study results, and an F-statistic of 11.13 with a p-value of 0.0025 on A-OKVQA user study results, indicating that the results are statistically significant.

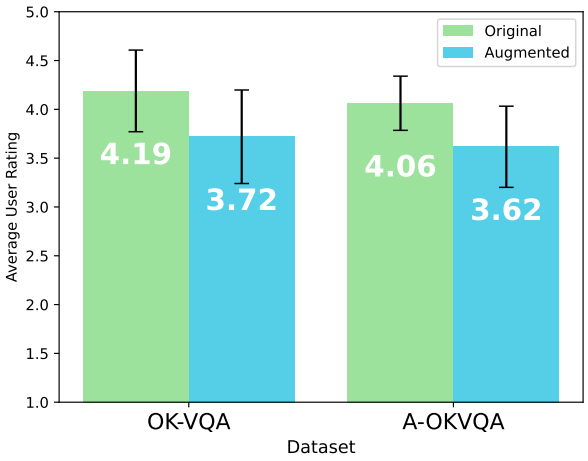

Figure 2. Results from the user study showing the average question relevance rating on OK-VQA and A-OKVQA datasets between the original and augmented image-question pairs obtained using our image-question matching strategy.

4.3. Ablation Study

Table 5 depicts an ablation study to find the significance of separating out rephrased questions from the diverse ones. These diverse questions add value to the retrieval training since they contain new information and hence can cover new scenarios in the image. Hence, we can observe a clear increase in performance when only diverse questions (with question similarity threshold $t \leq 0.8$) are used for augmentation instead of both rephrased and diverse questions on both the OK-VQA and A-OKVQA datasets. This confirms that the

additional augmentations, which are merely rephrased versions of originally annotated questions, do not add extra useful information.

Table 5. Comparison of the accuracy with and without the rephrased questions, controlled by the threshold t.

Dataset	Multiplicity	With Rephrased?	Threshold (t)	Accuracy
OK-VQA (evaluated on test set)	×2	✗	0.8	**55.74**
	×2	✓	-	55.62
	×3	✗	0.8	**56.27**
	×3	✓	-	55.97
A-OKVQA (evaluated on val set)	×2	✗	0.8	**54.96**
	×2	✓	-	54.64
	×3	✗	0.8	**55.48**
	×3	✓	-	55.28

4.4. Qualitative Visual Results

Figure 3 shows some visual examples that depict the newly matched questions for some images, which are obtained from SemIAug using BLIP retrieval. Note that in Figure 3, our image-question matching strategy, SemIAug, provides new question annotations that are relevant to the image. These questions can sometimes also cover objects that are not part of the original set of annotated questions, thus aiding in expanding the scene understanding while reducing the need for human annotators. Additional examples with BLIP-2 retrieval can be found in the Appendix A.

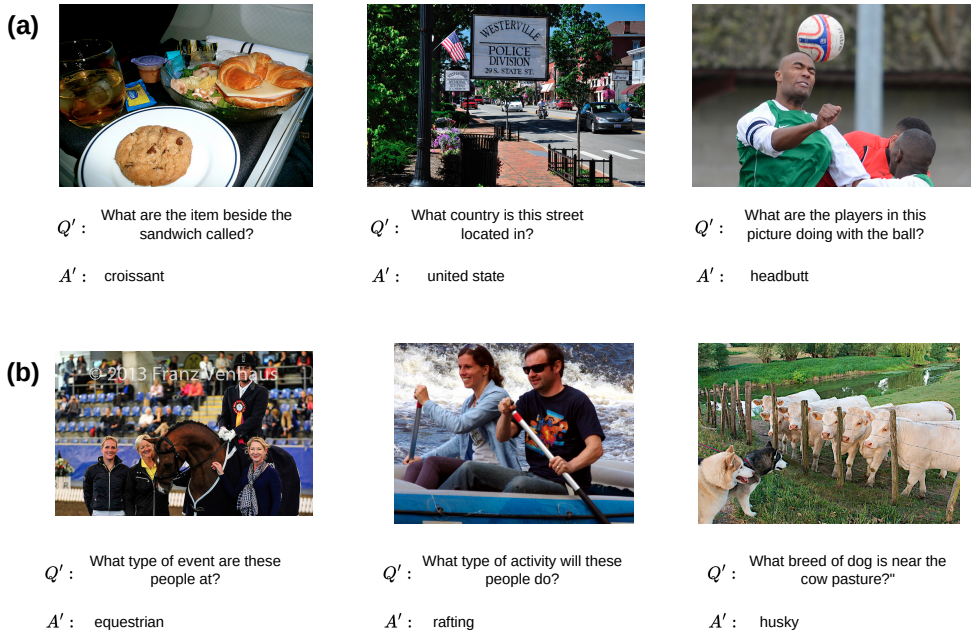

Figure 3. Qualitative visual results of BLIP retrieval with BLIP VQA. (**a**) Examples of newly matched questions $\{Q'\}$ and the corresponding answers $\{A'\}$ from the augmented OK-VQA dataset. (**b**) Examples of newly matched questions $\{Q'\}$ and the corresponding answers $\{A'\}$ from the augmented A-OKVQA dataset.

5. Discussion

In this work, we discuss a novel dataset augmentation technique, SemIAug, which utilises the implicit information in the dataset available in the form of unmatched but relevant image-question pairs. We draw parallels between this technique and semi-supervised learning approaches. We demonstrate the approach on data-scarce datasets OK-VQA and

A-OKVQA (which are also outside knowledge-based datasets) and show that the model performance can be enhanced without the need for any additional data/annotations or external knowledge sources. We show how the matched image-questions pairs are relevant and close to the original using a Qualitative user study. We compare the results of SemIAug with existing works that utilise additional data or require additional sophisticated models for data augmentation. We provide empirical arguments and ablation studies on how the different hyper-parameters were selected. SemIAug is a framework built around VLMs and their reusability. Hence, this technique can remain relevant in the future as it can adapt to any VLM and VQA dataset and will be usable as more powerful VLMs and more challenging datasets are developed.

6. Limitations and Future Work

An observation from the qualitative user study reveals that the additional annotations are not optimal but still reasonably pertinent to the image, while our augmentation strategy contributes significant performance enhancement, determining the optimal parameters for achieving the highest accuracy remains challenging due to the model dependency of both augmentation and the answering process.

Though our focus is on using SemIAug to produce new annotations for data-scarce VQA datasets, when applied to large datasets like VQAv2 [2], we did not see significant improvements in performance. This may be attributed to the fact that large datasets are already self-contained. Potential future directions would be to broaden the scope of SemIAug from the present model and data-agnostic setup to a task-agnostic setup.

Author Contributions: Conceptualization, methodology, software, formal analysis, data curation, writing—original draft preparation, visualization, B.D. and K.H.; writing—review and editing, supervision, A.N.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article.

Acknowledgments: Support from Institute of Eminence (IoE) project No. SB22231269EEETWO005001 for Research Centre in Computer Vision is gratefully acknowledged. Support from Microsoft India (R&D) Private Limited for the travel grant to attend the AAAI-24 workshop on Artificial Intelligence with Biased or Scarce Data (AIBSD) is gratefully acknowledged.

Conflicts of Interest: The author Bhargav Dodla is employed by Microsoft India (R&D) Private Limited. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as potential conflicts of interest.

Appendix A

This appendix provides additional details, results, and analyses of the SemiAug augmentation method proposed in the main paper. In Appendix A.1, we provide details of the dynamic multiplicity of question augmentation. In Appendix A.2, we experiment with the pretrained and finetuned models provided by the vision-language models (VLMs) owners. In Appendix A.3, we provide some more visual results of our augmentation on the OK-VQA and A-OKVQA datasets. In Appendix A.4, we provide some discussion and additional results of augmentation on sampled VQAv2 dataset, with our proposed method.

Appendix A.1. Dynamic Number of Question Augmentation

The number of questions annotated per image is limited in the case of data-scarce datasets, like OK-VQA and A-OKVQA. When additional annotations are incorporated into the model's training, there exists the potential for enhanced model performance in downstream tasks. However, it is not always guaranteed due to multiple reasons, such

as noisy annotations, rephrased questions, and the model's architecture and capability. As mentioned in the Section 3.1.3 in the main paper, we can set the number of potential questions matched for an image with hard or soft truncation. The soft truncation helps in matching the dynamic number of questions per image. The advantage of this is when many potentially good questions can be matched to an image, soft truncation allows all of the questions, unlike hard truncation. Similarly, it will not match a question just for the sake of matching if there is no additional good question match.

As we automatically annotate (using VLMs) more questions per image without the need for human efforts, there will be a problem of noisy annotations. It's very hard or impossible to eliminate the noisy annotations. However, with the dynamic number of question augmentation, we can attempt to reduce the possibility of noisy labels by selectively choosing the questions for an image.

After segregating rephrased and diverse questions as mentioned in the Section 3.1.2 of the main paper, the potential questions are sorted based on cosine similarity between images and those questions. We always consider originally annotated questions in the final augmented set as it is considered to be of good quality. The remaining questions are added to the original to obtain the augmented set. In the case of a fixed number of question matching, the additional required question for the image is taken from the potential questions set sorted based on cosine similarity. In the case of a dynamic number of augmentation, the number of questions to be considered is based on the predefined relevance threshold. The cosine similarities of the sorted questions are normalised, and only the questions whose normalised similarities are higher than the relevance threshold are considered for the further step. For the computational purpose, we set a lower and an upper limit for having a bound for the dynamic number of augmentation.

Appendix A.2. Pretrained and Finetuned Model Checkpoints

Recent advancements in the performance of vision-language tasks are due to improved vision-language pretraining. Large VLMs [13–17,35] use datasets with noisy image-text pairs collected from the web. BLIP [15] effectively utilises the noisy web data by bootstrapping the captions. Such pretrained models have very good zero-shot capabilities on downstream tasks like visual question answering (VQA), image-text matching, image captioning, video-text retrieval, etc. To assess the practical performance in downstream tasks, pretrained models are finetuned using task-specific datasets in advance, enabling their adaptation to the specific domain.

For SemIAug, as the starting point of model training or initialising weights, we considered the finetuned BLIP VQA checkpoint, in which the pretrained BLIP model was finetuned on VQAv2 [2], and visual-genome [29] datasets. Since BLIP has provided a finetuned VQA model on OK-VQA and A-OKVQA datasets, we used the same as the answering model for assessing the performance and did not see significant differences as compared to our finetuned versions of the same.

Appendix A.3. More Visual Results

Additional results on OK-VQA and A-OKVQA are provided in Figure A1. We can observe from the figure that the augmented questions are of decent quality, and the answers obtained using a pseudo-labelling approach are valid; however, the proposed augmentation is still noisy in many cases. With the human-in-the-loop, the bad augmentations can be removed, and we can obtain a good augmented dataset.

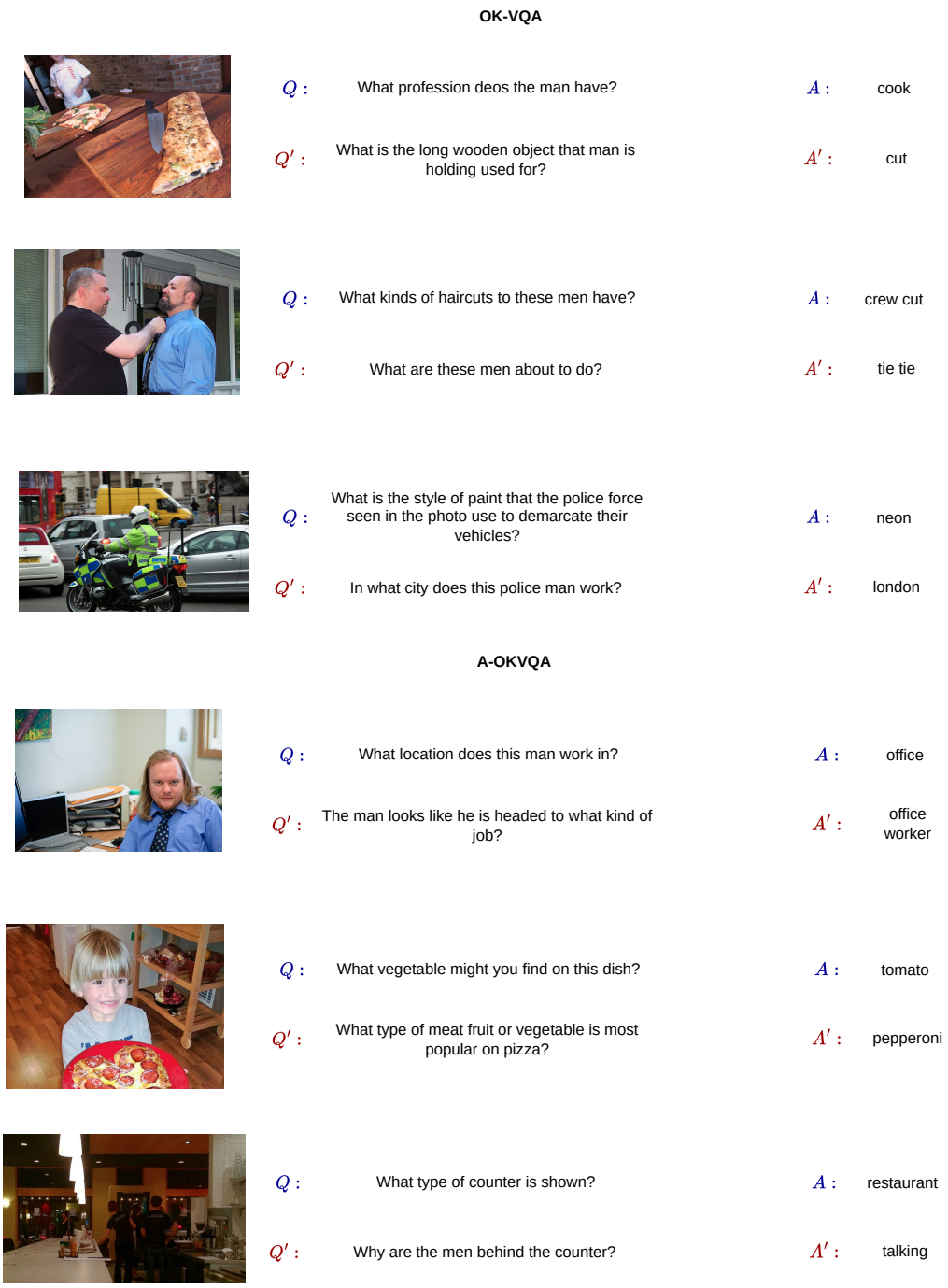

Figure A1. Examples of OK-VQA (top 3 rows) and A-OKVQA (bottom 3 rows), showing originally annotated question Q, answers A and newly matched question Q' and answer A' using SemiAug (using BLIP-2 as retrieval model and answered using BLIP VQA.

Appendix A.4. Results of Augmentation on VQAv2-Sampled Dataset

We conducted experiments involving our augmentation approach, SemIAug, using a subset of the VQAv2 dataset as our testing ground [2]. To mimic the behaviour of a data-scarce dataset, we randomly sampled 10% images (approximately 8000 images) from the VQAv2 dataset, and for each image, we chose two labelled annotations per image. Following this method, we created the *VQAv2-sampled* dataset, which resembles data-scarce datasets, for our experimentation purpose.

We experimented with the dataset by initialising the model with two different weights. Table A1 shows the set of experiments where the weights are initialised with a BLIP [15] pretrained checkpoint, and in Table A2 experiments, the model weights are initialised with the BLIP VQA checkpoint. We also conducted experiments where augmentation is performed with a fixed number of questions and with a dynamic number of questions per image. For the dynamic number of questions, we set the relevance threshold to 0.95, the lower limit to 2 and the upper limit to 5. The answering models for augmented questions are experimented with our semi-supervised approach and by directly utilising the BLIP VQA answering model.

We can see from Table A1 that the model initialised with BLIP pretrained model, finetuned on a small subset of VQAv2, yields poor performance with our augmentation strategy if answering is completed using our semi-supervised approach. This could be due to the fact that the model trained with a small amount of data is ineffective in answering the diverse questions from the VQAv2 validation set. However, if the newly augmented questions are answered using a finetuned BLIP VQA checkpoint, we can see good improvement in percentage gain, even though the absolute accuracy value is less. This indicates the effectiveness of augmentation with our Image-Question matching strategy.

We can observe a marginal improvement in accuracy when we initialise the model with a finetuned BLIP VQA checkpoint from Table A2. This is because of the strong weight initialisation and the amount of augmentation is much less compared to the data used for obtaining good finetuned weights used for initialisation. The table shows that with the dynamic number of good quality questions augmented, we can obtain similar or even better performances, sometimes with a lesser number of total questions for training. Instead of a sampled VQAv2 dataset, if we consider a full-scale VQAv2 dataset, the performance of the VLMs may be marginal or may not improve with the augmentation due to the fact that the dataset in itself has a sufficient amount of human labelled augmentations. With the use of automatically annotated noisy question-answers, there is a possibility of degradation in performance. Hence, augmentation is well suited for data-scarce datasets, and with the help of a good augmentation strategy, the performance of the VLMs can be improved by a good margin.

Table A1. Comparison of the performance of parameters using SemIAug, with the model initialized with a BLIP pretrained checkpoint. The model is evaluated on the VQAv2 validation (val) set.

Dataset	Type	Answering Setup	Ques. Augmentation	Ques Count	Multiplicity	Accuracy (%)	Gain (%)
VQAv2-sampled	original	-	-	16,000	×1	35.35	–
	augmented	semi-supervised	dynamic	23,817	×1.48	33.29	−2.06
	augmented	semi-supervised	fixed	32,000	×2	35.09	−0.26
	augmented	finetuned BLIP VQA	dynamic	23,817	×1.48	37.70	+2.35
	augmented	finetuned BLIP VQA	fixed	32,000	×2	39.12	+3.77

Table A2. Comparison of the performance of parameters using SemIAug, with the model initialized with a finetuned BLIP VQA checkpoint. The model is evaluated on the VQAv2 validation (val) set.

Dataset	Type	Answering Setup	Ques. Augmentation	Ques Count	Multiplicity	Accuracy (%)	Gain (%)
VQAv2-sampled	original	-	-	16,000	×1	66.41	–
	augmented	semi-supervised	dynamic	23,817	×1.48	66.58	+0.17
	augmented	semi-supervised	fixed	32,000	×2	66.53	+0.12
	augmented	finetuned BLIP VQA	dynamic	23,817	×1.48	66.62	+0.21
	augmented	finetuned BLIP VQA	fixed	32,000	×2	66.61	+0.2

References

1. Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C.L.; Parikh, D. Vqa: Visual question answering. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 2425–2433.
2. Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; Parikh, D. Making the vs. in vqa matter: Elevating the role of image understanding in visual question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6904–6913.
3. Agrawal, A.; Batra, D.; Parikh, D.; Kembhavi, A. Do not just assume; look and answer: Overcoming priors for visual question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4971–4980.
4. Marino, K.; Rastegari, M.; Farhadi, A.; Mottaghi, R. Ok-vqa: A visual question answering benchmark requiring external knowledge. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3195–3204.
5. Schwenk, D.; Khandelwal, A.; Clark, C.; Marino, K.; Mottaghi, R. A-okvqa: A benchmark for visual question answering using world knowledge. In Proceedings of the Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, 23–27 October 2022; Proceedings, Part VIII; pp. 146–162.
6. Garcia, N.; Ye, C.; Liu, Z.; Hu, Q.; Otani, M.; Chu, C.; Nakashima, Y.; Mitamura, T. A Dataset and Baselines for Visual Question Answering on Art. In Proceedings of the Computer Vision—ECCV 2020 Workshops, Glasgow, UK, 23–28 August 2020; pp. 92–108.
7. Lobry, S.; Marcos, D.; Murray, J.; Tuia, D. RSVQA: Visual question answering for remote sensing data. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 8555–8566. [CrossRef]
8. Hao, X.; Zhu, Y.; Appalaraju, S.; Zhang, A.; Zhang, W.; Li, B.; Li, M. MixGen: A New Multi-Modal Data Augmentation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops, Waikoloa, HI, USA, 3–7 January 2023; pp. 379–389.
9. Yang, A.; Miech, A.; Sivic, J.; Laptev, I.; Schmid, C. Just Ask: Learning to Answer Questions from Millions of Narrated Videos. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 1666–1677.
10. Grunde-McLaughlin, M.; Krishna, R.; Agrawala, M. AGQA: A Benchmark for Compositional Spatio-Temporal Reasoning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021.
11. Lu, J.; Batra, D.; Parikh, D.; Lee, S. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In Proceedings of the 33rd International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; Curran Associates Inc.: Red Hook, NY, USA, 2019; Article No. 2, pp. 13–23.
12. Li, L.; Yatskar, M.; Yin, D.; Hsieh, C.; Chang, K. VisualBERT: A Simple and Performant Baseline for Vision and Language. *arXiv* **2019**, arXiv:1908.03557.
13. Radford, A.; Kim, J.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning, Online, 18–24 July 2021; pp. 8748–8763.
14. Li, J.; Selvaraju, R.; Gotmare, A.; Joty, S.; Xiong, C.; Hoi, S. Align before fuse: Vision and language representation learning with momentum distillation. *Adv. Neural Inf. Process. Syst.* **2021**, *4*, 9694–9705.
15. Li, J.; Li, D.; Xiong, C.; Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In Proceedings of the International Conference on Machine Learning, Baltimore, ML, USA, 17–23 July 2022; pp. 12888–12900.
16. Li, J.; Li, D.; Savarese, S.; Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv* **2023**, arXiv:2301.12597.
17. Xu, X.; Wu, C.; Rosenman, S.; Lal, V.; Che, W.; Duan, N. Bridgetower: Building bridges between encoders in vision-language representation learning. *Proc. AAAI Conf. Artif. Intell.* **2023**, *37*, 10637–10647. [CrossRef]
18. Li, D.; Li, J.; Li, H.; Niebles, J.C.; Hoi, S.C.H. Align and prompt: Video-and-language pre-training with entity prompts. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 4953–4963.
19. Sharma, P.; Ding, N.; Goodman, S.; Soricut, R. 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), Melbourne, Australia, 15–20 July 2018; Association for Computational Linguistics: Stroudsburg, PA, USA, 2018; pp. 2556–2565.
20. Qi, D.; Su, L.; Song, J.; Cui, E.; Bharti, T.; Sacheti, A. Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data. *arXiv* **2020**, arXiv:2001.07966.
21. Lee, D. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. *Workshop Chall. Represent. Learn. ICML* **2013**, *3*, 896.
22. Rizve, M.N.; Duarte, K.; Rawat, Y.S.; Shah, M. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. *arXiv* **2021**, arXiv:2101.06329.
23. Cascante-Bonilla, P.; Tan, F.; Qi, Y.; Ordonez, V. Curriculum labeling: Revisiting pseudo-labeling for semi-supervised learning. *Proc. AAAI Conf. Artif. Intell.* **2021**, *35*, 6912–6920. [CrossRef]

24. Xu, Y.; Wei, F.; Sun, X.; Yang, C.; Shen, Y.; Dai, B.; Zhou, B.; Lin, S. Cross-model pseudo-labeling for semi-supervised action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 2959–2968.
25. Kil, J.; Zhang, C.; Xuan, D.; Chao, W.-L. Discovering the Unknown Knowns: Turning Implicit Knowledge in the Dataset into Explicit Training Examples for Visual Question Answering. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Punta Cana, Dominican Republic, 7–11 November 2021; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 6346–6361. [CrossRef]
26. Chen, L.; Zheng, Y.; Xiao, J. Rethinking Data Augmentation for Robust Visual Question Answering. In *Computer Vision–ECCV 2022*; Springer Nature: Cham, Switzerland, 2022; pp. 95–112. ISBN 978–3-031-20059-5.
27. Khan, Z.;Vijay K.; Schulter, S.; Yu, X.; Fu, Y.; Chandraker, M. Q: How To Specialize Large Vision-Language Models to Data-Scarce VQA Tasks? A: Self-Train on Unlabeled Images! In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 15005–15015.
28. Loper, E.; Bird, S. Nltk: The natural language toolkit. *arXiv* **2002**, arXiv:cs/0205028.
29. Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D.A.; et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.* **2017**, *123*, 32–73. [CrossRef]
30. Gui, L.; Wang, B.; Huang, Q.; Hauptmann, A.; Bisk, Y.; Gao, J. Kat: A knowledge augmented transformer for vision-and-language. *arXiv* **2021**, arXiv:2112.08614.
31. Lin, Y.; Xie, Y.; Chen, D.; Xu, Y.; Zhu, C.; Yuan, L. Revive: Regional visual representation matters in knowledge-based visual question answering. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 10560–10571.
32. Lu, J.; Clark, C.; Zellers, R.; Mottaghi, R.; Kembhavi, A. Unified-io: A unified model for vision, language, and multi-modal tasks. *arXiv* **2022**, arXiv:2206.08916.
33. Tan, H.; Bansal, M. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv* **2019**, arXiv:1908.07490.
34. Kamath, A.; Clark, C.; Gupta, T.; Kolve, E.; Hoiem, D.; Kembhavi, A. Webly supervised concept expansion for general purpose vision models. In *European Conference on Computer Vision*; Springer: Cham, Germany, 2022; pp. 662–681.
35. Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In Proceedings of the International Conference on Machine Learning, Online, 18–24 July 2021; pp. 4904–4916.