ArtQuest: Countering Hidden Language Biases in ArtVQA

Tibor Bleidt*
Hasso Plattner Institute
tibor.bleidt@hpi.de

Sedigheh Eslami*
Hasso Plattner Institute
sedigheh.eslami@hpi.de

Gerard de Melo
Hasso Plattner Institute
gerard.demelo@hpi.de

Abstract

The task of Visual Question Answering (VQA) has been studied extensively on general-domain real-world images. Transferring insights from general domain VQA to the art domain (ArtVQA) is non-trivial, as the latter requires models to identify abstract concepts, details of brushstrokes and styles of paintings in the visual data as well as possess background knowledge about art. This is exacerbated by the lack of high-quality datasets. In this work, we shed light on hidden linguistic biases in the AQUA dataset, which is the only publicly available benchmark dataset for ArtVQA. As a result, the majority of questions can be answered without consulting the visual information, making the "V" in ArtVQA rather insignificant. In order to counter this problem, we create a simple, yet practical dataset, ArtQuest, using structured information from the SemArt collection. Our dataset and the pipeline to reproduce our results are publicly available at https://github.com/bletib/artquest.

1. Introduction

The emergence of large foundation models has led to notable improvements in multimodal vision-language understanding tasks such as visual question answering (VQA; [8, 20, 32]). While these models have been extensively studied for general-domain tasks on generic real-world images, their capabilities in understanding specific domains such as art remains unclear. Art is a fundamental aspect of human culture, and art museums are visited by many millions of people every year. Thus, achieving visual question answering in the art domain (ArtVQA) is an important step towards conversational systems that can guide and assist people by addressing their information needs. Imagine encountering an interesting artwork and wondering who created it or in which time-frame it was created. ArtVQA can emit the answer to this, given a photo of the artwork and the relevant question in natural language. Furthermore, these systems may facilitate art education by acting as a study assistant.

Achieving ArtVQA is a challenging task, since the model needs to understand the detailed visual information in paintings, e.g., brushstrokes and common patterns in artistic styles for inferring information about the artist, type or art movements from the painting. This visual information is also often represented at different levels of abstraction, making the visual understanding quite different from the understanding of general-domain images. Moreover, the model needs to interpret the natural language question and associate it with the visual data. ArtVQA also requires the model to possess background knowledge about the historical context of artworks, e.g., "when was the painting created?" [12].

Our work employs a generative approach for ArtVQA using a prefix language modeling objective. We investigate AQUA, the only publicly available benchmark dataset for ArtVQA and identify hidden language biases that exist in this data, casting doubt on its value for VQA evaluation. In particular, we show that, due to hidden biases, the majority of questions can be answered without any dependency on the visual information, making the "V" in VQA rather insignificant. These biases can falsely suggest that AI models are making progress in visual understanding of artworks. This observation motivated us to provide a cleaner, more reliable, and less biased dataset for the task of ArtVQA that genuinely requires consulting the visual data to answer knowledge-seeking questions. We propose ArtQuest (**Art Quest**ions), a new set of question–answer pairs for the paintings in the SemArt collection [11] using the structured information in this collection. We show that ArtQuest elevates the importance of visual data for answering questions and hence, allows for a more reliable training and evaluation of ArtVQA models. While ArtQuest consists of simple types of questions, we believe it is the first step for enabling reliable benchmarking of ArtVQA models. To the best of our knowledge, this is the first work to study linguistic biases in ArtVQA as well as to evaluate the performance of state-of-the-art vision and language models in the art domain.

*equal contribution

2. Related Work

Visual Question Answering. Several attempts aim at solving VQA as a classification task for predicting the unique answers seen in the training dataset [2, 18]. Recent research shows rapid progress in VQA using Vision-Language Pre-training (VLP). VLP learns effective representations for both visual and textual data while capturing their correspondence. Existing VLP approaches use a large amount of image–text pairs and pre-train tasks such as contrastive learning of vision-language data [13,27,38], prefix language modeling [37], image-conditioned masked language modeling or text-conditioned masked image modeling [4, 15, 34] as well as image-conditioned causal language modeling [20, 21, 39]. Most of these VLP models employ a fully-connected layer on top of their VLP architecture to recast the VQA task as classification [4, 10, 32, 36]. [1] implemented generative versions of ViLBERT [24] and ALBEF [22] and showed that generative approaches tend to result in better out-of-distribution generalization. Inspired by this, we choose a generative prefix language modeling approach for solving ArtVQA. Regarding language priors and biases in the general domain VQA, authors in [14] published a new benchmark dataset with reduced language priors. In [29], adversarial regularization by training with question-only adversary setting has been proposed. Furthermore, [29] aims at reducing the superficial correlations between questions and their corresponding frequent answers by adding the objective of distinguishing superficial similar instances in the training step.

Vision-Language and VQA for Art. In the art domain, VLP has enabled recent advances in artistic image generation from text prompts [5, 30, 31]. In CLIP-Art [7], the contrastive vision-language loss from CLIP [27] is used to fine-tune on the iMet collection [40], leading to improvements on downstream multimodal retrieval and classification tasks for paintings. [3] present a framework for generating informative painting captions based on masked sentence generation using LSTM and knowledge retrieval using TF-IDF vectors. Authors report their experimental results on the SemArt collection [11]. For ArtVQA, [11] made notable contributions by introducing the AQUA dataset and the VIKING baseline. The knowledge question–answer pairs in AQUA were generated using rule-based approaches similar to [16]. For visual questions, the authors employed two different approaches. One was to use iQAN [23] to generate questions along with Amazon Rekognition for object detection as well as answer generation. Another approach was to use Pythia [33] to generate captions for each painting and then apply rule-based approaches on the generated captions so as to obtain question–answer pairs. Our work provides detailed analyses that reveal linguistic biases in AQUA. Subsequently, we propose ArtQuest in order to avoid such biases.

3. Prefix Language Modeling for ArtVQA

In this work, VQA is formulated as a generative sequence-to-sequence modeling task with the objective of Prefix Language Modeling (PrefixLM) [37]. For a sequence $\mathbf{s}$, the goal of PrefixLM is to auto-regressively predict $\mathbf{s}_{\setminus \mathbf{t}}$ conditioned on the prefix sequence $\mathbf{s}_{\mathbf{t}}$, where $\mathbf{s}_{\mathbf{t}} \oplus \mathbf{s}_{\setminus \mathbf{t}} = \mathbf{s}$. The symbol $\oplus$ is used to denote concatenation throughout this work.

Closed-book ArtVQA. In the closed-book VQA scenario, given a training dataset $\mathcal{T} = \{(v_i, q_i, a_i)\}_{i=1}^{D}$ of size D, where v_i is a painting, q_i is an associated natural language question, and a_i is the corresponding answer in natural language, our goal is to train a model that generates the correct answer a_i given image–question pair (v_i, q_i). We assume encoding functions to obtain $\mathbf{f}_v \in \mathbb{R}^n$ as an n-dimensional vector encoding for image v_i and the text embedding $\mathbf{f}_q \in \mathbb{R}^{m \times l}$ for the question q_i with l tokens. We apply a fully connected projection layer $F : \mathbb{R}^n \to \mathbb{R}^m$ on f_v and obtain $\mathbf{f'}_v \in \mathbb{R}^m$, in order to unify embedding sizes of the encoded image and question. Furthermore, a transformer-based encoder–decoder architecture is used to achieve PrefixLM, where the decoder receives $\mathbf{f}_{vq} = [\mathbf{f'}_v \oplus \mathbf{f}_q] \in \mathbb{R}^{m \times (l+1)}$. The decoder is then trained to generate the encoded answer $\mathbf{f}_a \in \mathbb{R}^{m \times l'}$ for the answer a_i with l' tokens. We generate each token in the answer sequence auto-regressively with cross-entropy as the loss function. In this approach, the concatenation of the encoded visual data and the cross-attention in the transformer decoder enables the VQA model to incorporate visual information for answer generation.

Open-book ArtVQA. We also consider ArtVQA in an open-book scenario, where the model is allowed to see an additional explanatory caption c_i about the painting v_i when providing the answer a_i to the question q_i. The motivation for this is that in ArtVQA, answering a question might require external background information not explicitly present in the painting. Therefore, the model may use the additional information in the explanatory text to elicit the correct answer to the question.

We employ an image–text retrieval approach in order to fetch the most relevant caption c_i for a query image v_i from a database containing art-related captions C. Using the appropriate encoding functions, each caption c with l'' tokens is encoded to obtain $\mathbf{f}_c \in \mathbb{R}^{m \times l''}$. We then employ average pooling to achieve a `[CLS]`-level embedding for representation of the caption sequence as $\mathbf{f'}_c \in \mathbb{R}^m$. The image v_i is also encoded as $\mathbf{f}_v \in \mathbb{R}^n$. Similar to the previous section, we apply a fully connected layer to unify the embedding dimensions. We then apply 1-Nearest Neighbor with cosine similarity to identify the closest caption in the database for

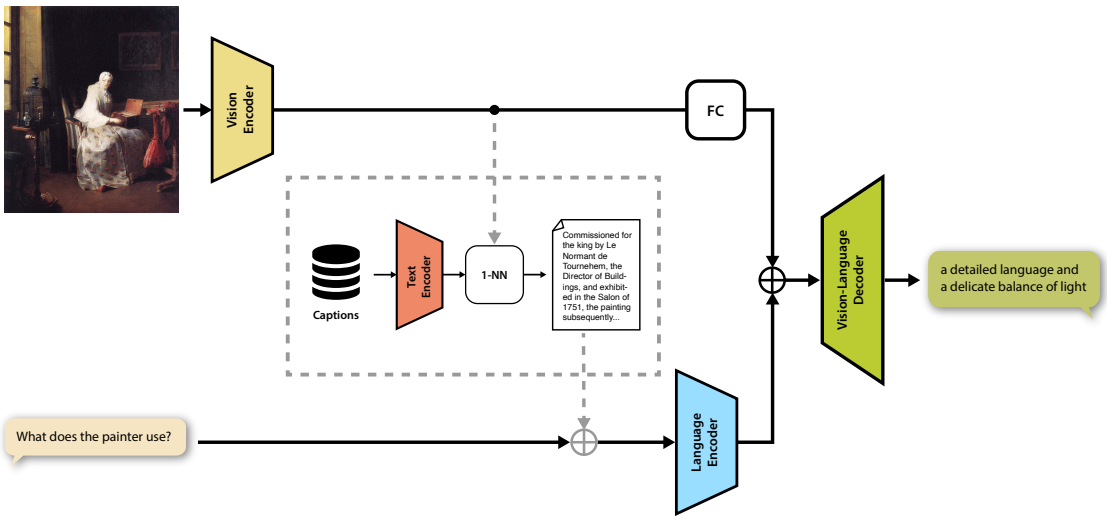

Figure 1. Schematic architecture of our PrefixLM model. Gray parts are only active in the open-book VQA setting.

the query image v_i:

$$c_i = \underset{c \in C}{\operatorname{argmax}} \frac{\mathbf{f}_v^\intercal \, \mathbf{f}_c}{|\mathbf{f}_v| \, |\mathbf{f}_c|}. \tag{1}$$

The retrieved caption c_i is then concatenated to the question following the template:

$$T_i = \text{“question: } \{q_i\} \text{ context: } \{c_i\}\text{”}. \tag{2}$$

as suggested by [28].

Using the text encoding function, embedding $\mathbf{f}_t \in \mathbb{R}^{m \times l}$ is achieved for the final text T_i. For answer generation by PrefixLM, the decoder receives $\mathbf{f}_{vt} = [\mathbf{f'}_v \oplus \mathbf{f}_t] \in \mathbb{R}^{m \times (l+1)}$. The decoder is then optimized for auto-regressive generation of the encoded answer $\mathbf{f}_a \in \mathbb{R}^{m \times l'}$ for the answer a_i with l' tokens using the cross-entropy loss. Using the cross-attention in the transformer decoder which receives the concatenated encoded visual data, answer generation is grounded on the visual information in our approach.

A schematic overview of our approach is shown in Fig. 1. In the open-book scenario, the optional retrieval module is activated and retrieves the most relevant caption for the given image.

4. Uncovering Language Biases in AQUA

Underlying language biases in VQA datasets can lead to the false impression that VQA models are making progress towards truly understanding images when they merely exploit language priors to achieve a high accuracy. Inspired by [14], we started our study on ArtVQA with the goal of understanding whether the “V” truly matters in ArtVQA. We used the only available benchmark dataset, AQUA, which encompasses two kinds of questions, visual and knowledge ones, which we each evaluated for language biases.

Visual questions. Authors in [12] define visual questions as questions that mainly target visual contents in paintings, e.g, “what do a group of men stand next to?”.[1] For studying language biases in visual questions, we used the solution illustrated in Section Sec. 3, while ignoring the image input. We considered the closed-book VQA scenario and used BART-base [19] as well as Flan-T5-small [6] for the PrefixLM modeling. In this case, for example, when using BART-base, we used BART’s encoder as our Language Encoder and BART’s decoder as the Vision-Language Decoder. The same setup was repeated when using the T5 model.

We performed our experiments in zero-shot as well as fine-tuned settings. Our motivation for reporting the zero-shot results of BART and T5 is to show that these models do not have prior background knowledge with respect to the art domain. When fine-tuning, we used question–answer pairs and fine-tuned the encoder–decoder model end-to-end. The encoder received questions and the decoder generated answers auto-regressively. We set the maximum length for the encoder input to 512 and for the decoder output to 100 tokens. Our final evaluations are performed on the AQUA test split. Results of our experiments are given in Tab. 1.

As can be seen, the zero-shot results of BART and T5 are quite weak on ArtVQA, confirming the assumption that these models do not already possess much prior knowledge in the art domain. However, once these models are fine-tuned merely on the question–answer pairs from AQUA without the presence of images, they reach up to 71% accuracy on answering visual questions. Achieving 71% accuracy with no visual data strongly points to the presence of hidden language biases in the visual questions, making the “V” fairly insignificant for this dataset.

[1] This example is taken directly from the AQUA train set.

Visual Encoder	Language Encoder-Decoder		Exact Match Accuracy
None	BART	Zero-shot	0.5%
		Fine-tuned	71.0%
CLIP-ViT-B/32	BART	Fine-tuned	78.3%
None	Flan-T5	Zero-shot	0.2%
		Fine-tuned	70.9%
CLIP-ViT-B/32	Flan-T5	Fine-tuned	79.5%

Table 1. Accuracy of **closed-book** VQA on **visual** questions from AQUA. Yellow highlighting corresponds to experiments without images (Visual Encoder: None), while cyan denotes experiments with images.

In order to further evaluate the effectiveness of our PrefixLM model, we repeated our experiment while considering the images. The results of this are also reported in Tab. 1. We use the CLIP visual encoder [27] with ViT-B/32 back-end for encoding paintings. Our results show that using the visual data results in about absolute 8% improvement in the accuracy of answering visual questions. This observation evinces that our PrefixLM modeling is effective in incorporating the visual information for the task of VQA.

Knowledge questions. [12] describes knowledge questions as questions that require background knowledge in the art domain for answering. These questions have been created by applying rule-based approaches such as those by [16] on the descriptions from the SemArt dataset [11]. The fact that the visual data has not been considered in the process of creating question–answer pairs is our first clue regarding whether "V" really matters for answering these questions.

We performed closed-book as well as open-book VQA while ignoring the input images. We again employed our PrefixLM encoder–decoder approach from Sec. 3. For the closed-book scenario, we repeated the kinds of experiments described above for visual questions using BART-base and Flan-T5-small. The results are summarized in Tab. 2, where we report the accuracy scores. In this setting, the poor performance of both BART and T5 at answering knowledge questions in the closed-book scenario makes it apparent that achieving closed-book VQA for knowledge questions is a challenging task. Even after fine-tuning, these models achieve an accuracy of less than 13%.

In the open-book scenario, in order to assess the performance without the presence of visual data, we adapted our retrieval module in Fig. 1 to work without the image input. For this, we considered two approaches:

1. Question-based Caption Retrieval (QCR) using TF-IDF vectors and cosine similarity. We pick the top-10 captions and re-rank them by training a BERT-base classifier [9]. The classifier receives a question and one of the top-10 captions c and learns a binary classification $F : (q_i, c) \rightarrow \{0, 1\}$. This approach is inspired by [12].

2. Oracle Method (OM) of fetching the corresponding caption for each image from the SemArt collection. Here, we use image names for finding the overlap between AQUA and the SemArt datasets.

For the open-book scenario, we experimented with the Flan-T5-small model. We chose T5, since it has been already optimized for supporting open-book question answering. The encoder received the concatenated question and caption using the template described in Eq. (2). The maximum length for the encoder input and decoder output are set to 512 and 100 tokens, respectively.

Results of our experiments in Tab. 3 illustrate that fine-tuning Flan-T5 using the QCR and OM retrieval approaches enables answering knowledge questions with an accuracy of up to 77.6% and 85%, respectively. These high scores alone are not necessarily an indicator of language bias in knowledge questions, since it could be that the captions already provide informative details about what is present in the visual data. However, we observe that when adding the visual component by encoding images and using PrefixLM, the accuracy stays the same, showing that the visual data is not playing an essential role for comprehending and answering questions. This observation once again suggests that "V" plays a negligible role for VQA in the AQUA dataset.

In addition, we provide qualitative examples in Fig. 2 to illustrate how knowledge questions do not depend on the visual information and can be answered regardless of the image. Based on OK-VQA [25], we believe it is necessary for knowledge questions in a VQA task to include dependencies and references to the visual information. As an example, OK-VQA includes the question "what sort of vehicle uses *this item*?", which is asked about an image from a fire hydrant in a street. The ground truth answer to this question is "firetruck". In this example, due to the mention of "this item" in the question, the VQA model must gain insights about the objects in the image, connect the question and objects, and ultimately, determine the correct answer.

5. ArtQuest: Art "Quest"ions

Question-answer generation. Given the shortcomings observed for the sole available benchmark dataset for ArtVQA, there is an urgent need to curate a more reliable, less biased ArtVQA benchmark. To this end, we harnessed the paintings and the structured information from the SemArt dataset [11]. We used the artist, title, technique, school, time-frame, and type attributes from SemArt and manually created six initial open-ended English language questions for each painting.

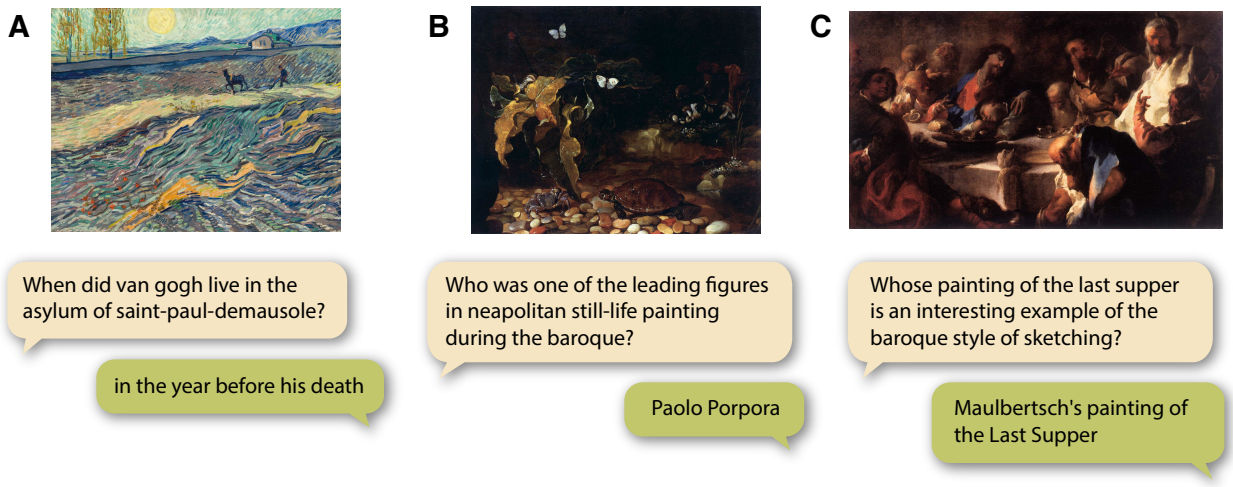

Figure 2. Knowledge question examples from AQUA where the question–answer pair is independent of the image.

Visual Encoder	Language Model			Exact Match Accuracy
None	BART	Closed-book	Zero-shot	0.1%
			Fine-tuned	6.0%
CLIP ViT-B/32	BART	Closed-book	Fine-tuned	5.3%
None	Flan-T5	Closed-book	Zero-shot	0%
			Fine-tuned	12.1%
CLIP ViT-B/32	Flan-T5	Closed-book	Fine-tuned	12.9%

Table 2. Accuracy of **closed-book** VQA on **knowledge** questions from AQUA. Yellow highlighting corresponds to experiments without images (Visual Encoder: None), while cyan denotes experiments with images.

Visual Encoder	Language Model			Accuracy
None	Flan-T5	Open-book-QCR	Zero-shot	21.3%
			Fine-tuned	77.6%
		Open-book-OM	Fine-tuned	85%
CLIP ViT-B/32	Flan-T5	Open-book-QCR	Fine-tuned	77.6%
		Open-book-OM	Fine-tuned	84.9%

Table 3. Accuracy of **open-book** VQA on **knowledge** questions from AQUA. Yellow highlighting corresponds to experiments without images (Visual Encoder: None), while cyan denotes experiments with images.

These initial questions are denoted by:

$$
\begin{aligned}
Q = \{&\text{“Who is the artist of this image?”},\\
&\text{“What is the title of this painting?”},\\
&\text{“What painting technique is used?”},\\
&\text{“What is the school of the painting?”},\\
&\text{“In which time-frame was the painting painted?”},\\
&\text{“What is the type of this painting?”}\}
\end{aligned}
$$

We consider these questions to represent knowledge questions whose answers depend on the visual content of the paintings. Despite these questions being fairly simple, we argue that achieving ArtVQA models that answer these questions correctly is the first step towards developing reliable ArtVQA systems. In future work, we plan to expand the diversity of the questions with new versions of ArtQuest.

In order to engender greater lexical variety, we employed ChatGPT [26] to rephrase our initial questions. We used the prompt:

> *“Rephrase each of the following questions 5 times. Be as short and precise as possible!”*

We invoked this prompt two times and manually selected a combination of the best generated questions from a human’s perspective.As a result, we ended up having 5 differently phrased versions for each initial question[2] in Q:

$$Q' = \{Q_i\}_{i=1}^{6} \quad \text{where} \quad Q_i = \{q_i^j\}_{j=1}^{5}$$

For each painting, among the 5 question versions in Q_i for each type, we randomly selected one version to use with that painting. As a result, we have 6 questions per painting,

[2]This includes the initial versions.

which are expressed in somewhat different ways across the paintings. As shown in Fig. 4, our question creation approach ensures a balanced question type distribution in ArtQuest.

Questions sharing the same semantics across the paintings in ArtQuest is beneficial for making the dataset less prone to linguistic biases. According to [14], one approach for reducing language priors is to ensure that given a triplet (I, Q, A) of image, question, and answer, there exists an I' such that the answer to Q is $A' \neq A$. In ArtQuest, since the questions are semantically shared across all paintings, this condition is already satisfied. For example, a question like "Who is the artist of this image?" is asked about each painting and therefore is intrinsically less likely to always be answered with the same painter in the ArtQuest dataset. This is also supported by Fig. 4, which shows that ArtQuest includes works from a large number of artists. The same reasoning holds for other types of questions in ArtQuest.

The answer for each generated question is taken from the corresponding attribute value in the SemArt collection. For example, for the question "Who is the *artist* of this image?", we consider the painting name to match the painting in the SemArt dataset and then retrieve the value of the "artist" column in SemArt as the ground truth answer for the question.

Finally, we randomly selected 100 paintings from ArtQuest and asked an art expert to manually review the correctness of the 600 created question–answer pairs. All of the generated question–answer pairs were annotated as correct. In Fig. 3, a qualitative example of our created question–answer pairs can be seen.

Dataset analysis. We followed the splits in the SemArt [11] dataset also for the generated question–answer pairs. Thus, there are more than $17K$, $1K$, and $1K$ unique paintings in the training, validation, and test sets, respectively. For each image, we created the 6 different question–answer pairs as described before.

We present the answer distribution for each question type in Fig. 4. The majority of questions about school, technique, time-frame, and type are answered by Italian, oil on canvas, 1601–1650, and religious, respectively. The corresponding ZeroR baselines for Italian as school, oil on canvas as technique, 1601–1650 as time-frame and religious as type are 41.4%, 47.1%, 17.7% and 38.8%, respectively. As can be seen, the answer distribution still makes it hard to obtain a very high accuracy when merely relying on priors. Moreover, in the distributions for the artist and title questions, there is a very wide variety of answer values.
The distribution of question lengths in ArtQuest is plotted in Fig. 5. We observe that the majority of questions in ArtQuest include 5–7 words. Finally, we provide the distribution of questions by their first four words in Fig. 6.

Testing for language biases. We repeated the experiments from Sec. 4 once again but here evaluated whether there are language biases in ArtQuest. We used the BART-base model as well as the Flan-T5-small with and without the presence of images. The results are given in Tab. 4. We observe that zero-shot BART and T5 achieve very low accuracy scores, once again, showing that these models do not carry sufficient prior art knowledge. Fine-tuning BART without the presence of images achieves about 20%. This increase is due to the imbalance of the answer distribution in ArtQuest. As shown in Fig. 4, e.g., the majority of paintings in the dataset are from the Italian school. Therefore, a question such as "What is the school of the painting?" may get biased towards always answering "Italian". However, the overall language bias is found to be small. Without the visual information, the model cannot achieve a very high accuracy on ArtQuest. The same trend of explanation applies when fine-tuning closed-book and open-book T5 without images. In contrast, when testing with the presence of images, we observe substantial improvements of up to around 30% in the model's ability to correctly answer closed-book questions. This shows that the presence of "V" is significant for answering questions in ArtQuest. In the open-book scenario, we observe that answering the questions when only using the captions reaches up to 63% accuracy. This is because the SemArt captions contain background information about the painting and can include information such as title, artist, etc. We also test the open-book scenario with images and observe that using images for VQA improves the accuracy by up to 3%. This observation concludes that visual information provides additional information for the VQA model.

Visual Encoder	**Language Model**			**Exact Match** Accuracy
None	BART	Closed-book	Zero-shot	0%
			Fine-tuned	20.4%
CLIP ViT-B/32	BART	Closed-book	Fine-tuned	36.9%
None	Flan-T5	Closed-book	Zero-shot	0%
			Fine-tuned	23.4%
CLIP ViT-B/32	Flan-T5	Closed-book	Fine-tuned	50.2%
None	Flan-T5	Open-book-QCR	Zero-shot	3.5%
			Fine-tuned	13%
		Open-book-OM	Fine-tuned	63.6%
CLIP ViT-B/32	Flan-T5	Open-book-QCR	Fine-tuned	22.8%
		Open-book-OM	Fine-tuned	66.6%

Table 4. Accuracy of open-book and closed-book VQA on ArtQuest. Yellow highlighting corresponds to experiments without images (Visual Encoder: None), while cyan denotes experiments with images.

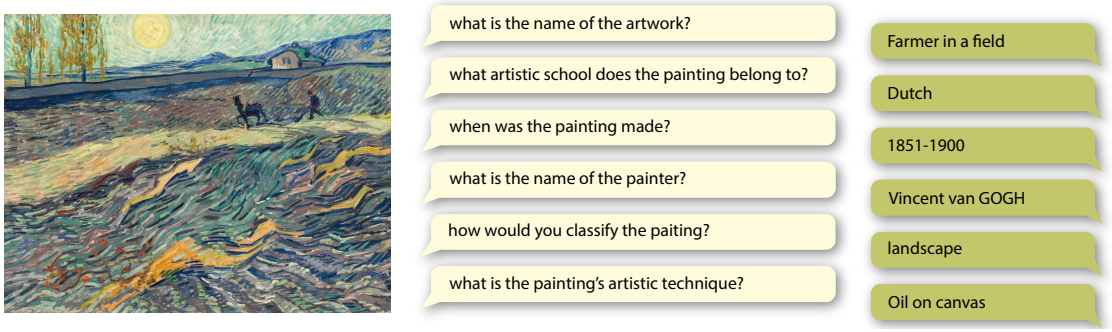

Figure 3. Examples taken directly from ArtQuest.

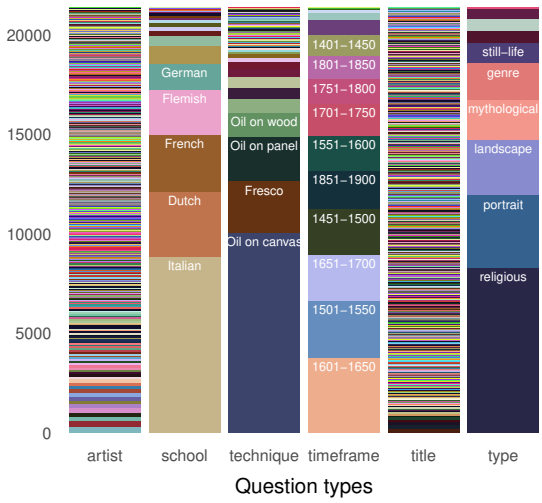

Figure 4. Distribution of answers per question type.

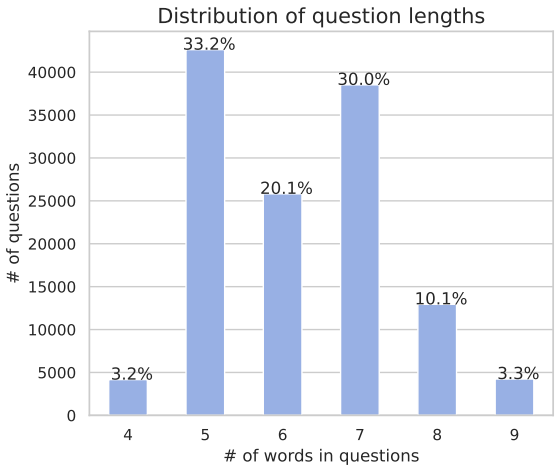

Figure 5. Distribution of question lengths in ArtQuest.

VQA Model		Accuracy	BLEU
OFA-base (ZS)	Closed-book	1.9%	4%
BLIP-base (ZS)	Closed-book	2.4%	5.1%
VIKING* (FT)	Closed-book	37.9%	39.5%
	Closed-book	**50.2%**	**56.3%**
PrefixLM (FT)	Open-book	**53.5%**	**59.8%**

Table 5. Results of zero-shot (ZS) and fined-tuned (FT) VQA baselines on ArtQuest. * is the model from [12].

6. Benchmarking VQA Models

In this section, we provide baselines for VQA when using ArtQuest. For zero-shot closed-book VQA, we considered OFA [35] and BLIP [21], which are amongst the top VQA models in the general domain VQA leaderboard[3]. Our experimental results in Tab. 5 show that these models achieve less than 5% accuracy and BLEU score on ArtQuest. This shows that general-domain vision-language models lack prior art knowledge and do not generalize to the specific art domain.

When fine-tuning, we tested the VIKING closed-book ArtVQA model from [12] as well as our proposed PrefixLM model in both closed-book and open-book settings. VIKING was trained for 10 epochs with batch size 512. VIKING employs LSTM for encoding questions, ResNet-152 encoding paintings and Bilinear Attention Networks [18] for fusing the encoded questions and paintings.

In the PrefixLM model, we used CLIP ViT-B/32 as the visual encoder, the encoder from Flan-T5-small as the language encoder, and the decoder from Flan-T5-small as the vision-language decoder. We set the maximum sequence length for the language encoder and vision-language decoder to 512 and 100, respectively. For the retrieval module, we used the text encoder from CLIP ViT-B/32 to encode the captions

[3]As of August 2023: https://eval.ai/web/challenges/challenge-page/830/leaderboard/2278

Figure 6. Distribution of questions by their first four words.

	artist	**title**	**technique**	**school**	**time-frame**	**type**
	EM |BLEU	EM |BLEU	EM |BLEU	EM |BLEU	EM |BLEU	EM |BLEU
Closed-book	22.6% |24.2%	8% |16.8%	60.8% |75.8%	68% |68%	60.1% |71%	81.8% |81.8%
Open-book	34.1% |36.4%	11.5% |22.9%	61.2% |75.1%	69.4% |69.4%	61.5% |72.1%	83.2% |83.2%

Table 6. Acccuracy and BLEU scores of VQA for each question type when using our PrefixLM. EM stands for Exact Match accuracy.

from SemArt and truncated the captions to 76 tokens. Subsequently, 1-Nearest Neighbor with cosine similarity was used to select the most relevant caption per painting. For the implementation, we used the Faiss library [17]. The top@1 accuracy for our retrieval module was 46%.

The overall results of the fine-tuned models are provided in Tab. 5. The VIKING model achieves about 38% accuracy with a multi-label classification approach, cluing that treating ArtQuest with VIKING's classification approach is not effective. In contrast, the generative PrefixLM model achieves a better baseline of 50% and 53.5% exact match accuracy in closed-book and open-book settings, respectively. We hypothesise that by improving the accuracy of the retrieval module, stronger baselines can be achieved. Furthremore, in Tab. 6, detailed accuracy and BLEU scores of using the PrefixLM model at answering each question type is provided. We observe that answering questions about the type of the painting is easier in comparison to the other question types in ArtQuest. Furthermore, correctly answering with the titles of artworks appears to be a challenging task. This is because the paintings in the test set are unseen in training and validation sets and hence not specifically get learned during training. The open-book setting also does not achieve a great accuracy, since our top@1 retrieval performance of the retrieval module is only 46%. It is also apparent that predicting the artist is another challenging task. As shown in Fig. 4, for many of the artists in ArtQuest, there exists very few paintings. Therefore, few-shot learning approaches might be required to predict the artist more effectively. We hope that the baselines provided in this work motivate researchers to conduct further research on enhancing ArtVQA.

Conclusion

This work provides an extensive study on the current state of VQA in the art domain. We show that the only previously available benchmark dataset for ArtVQA is biased towards language priors, and hence, does not require considering the input image for answering questions. In order to address this problem, we propose ArtQuest as a new benchmark dataset for ArtVQA and through extensive experiments, show that ArtQuest does not suffer from language biases.

Acknowledgements. We acknowledge the financial support by the German Federal Ministry for Education and Research (BMBF) within the project KI-Servicezentrum Berlin Brandenburg (01IS22092) as well as the HPI-MIT Designing for Sustainability program.

References

[1] Aishwarya Agrawal, Ivana Kajić, Emanuele Bugliarello, Elnaz Davoodi, Anita Gergely, Phil Blunsom, and Aida Nematzadeh. Rethinking evaluation practices in visual question answering: A case study on out-of-distribution generalization. *arXiv preprint arXiv:2205.12191*, 2022. 2

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

[3] Zechen Bai, Yuta Nakashima, and Noa Garcia. Explain me the painting: Multi-topic knowledgeable art description generation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5402–5412. IEEE, 2021. 2

[4] Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *Advances in Neural Information Processing Systems*, 35:32897–32912, 2022. 2

[5] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023. 2

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

[7] Marcos V Conde and Kerem Turgutlu. Clip-art: Contrastive pre-training for fine-grained art classification. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3951–3955. IEEE Computer Society, 2021. 2

[8] Wenliang Dai, Lu Hou, Lifeng Shang, Xin Jiang, Qun Liu, and Pascale Fung. Enabling multimodal generation on clip via vision-language knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2383–2395, 2022. 1

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 4

[10] Sedigheh Eslami, Christoph Meinel, and Gerard De Melo. Pubmedclip: How much does clip benefit visual question answering in the medical domain? In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1151–1163, 2023. 2

[11] Noa Garcia and George Vogiatzis. How to read paintings: semantic art understanding with multi-modal retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 1, 2, 4, 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 *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 92–108. Springer, 2020. 1, 3, 4, 7

[13] Shashank Goel, Hritik Bansal, Sumit Bhatia, Ryan Rossi, Vishwa Vinay, and Aditya Grover. Cyclip: Cyclic contrastive language-image pretraining. *Advances in Neural Information Processing Systems*, 35:6704–6719, 2022. 2

[14] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 2, 3, 6

[15] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 2

[16] Michael Heilman and Noah A Smith. Good question! statistical ranking for question generation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 609–617, 2010. 2, 4

[17] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019. 8

[18] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. *Advances in neural information processing systems*, 31, 2018. 2, 7

[19] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, 2020. 3

[20] 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 preprint arXiv:2301.12597*, 2023. 1, 2

[21] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 2, 7

[22] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 2

[23] Yikang Li, Nan Duan, Bolei Zhou, Xiao Chu, Wanli Ouyang, Xiaogang Wang, and Ming Zhou. Visual question generation as dual task of visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6116–6124, 2018. 2

[24] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 2

[25] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204, 2019. 4

[26] OpenAI. Chatgpt (june 15 version). https://chat.openai.com, 2023. 5

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 4

[28] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020. 3

[29] Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. Overcoming language priors in visual question answering with adversarial regularization. *Advances in Neural Information Processing Systems*, 31, 2018. 2

[30] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 2

[31] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 2

[32] Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. How much can CLIP benefit vision-and-language tasks? In *International Conference on Learning Representations*, 2022. 1, 2

[33] Amanpreet Singh, Vedanuj Goswami, Vivek Natarajan, Yu Jiang, Xinlei Chen, Meet Shah, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Mmf: A multimodal framework for vision and language research. https://github.com/facebookresearch/mmf, 2020. 2

[34] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*, 2020. 2

[35] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *arXiv preprint arXiv:2202.03052*, 2022. 7

[36] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19175–19186, 2023. 2

[37] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021. 2

[38] Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. Wav2clip: Learning robust audio representations from clip. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4563–4567. IEEE, 2022. 2

[39] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 2

[40] Chenyang Zhang, Christine Kaeser-Chen, Grace Vesom, Jennie Choi, Maria Kessler, and Serge Belongie. The imet collection 2019 challenge dataset. *arXiv preprint arXiv:1906.00901*, 2019. 2