electronics

Article

A Picture May Be Worth a Hundred Words for Visual Question Answering †

Yusuke Hirota [1,*], Noa Garcia [1], Mayu Otani [2], Chenhui Chu [3] and Yuta Nakashima [1]

[1] Graduate School of Information Science and Technology, Osaka University, Osaka 565-0871, Japan; noagarcia@ids.osaka-u.ac.jp (N.G.); n-yuta@ids.osaka-u.ac.jp (Y.N.)
[2] CyberAgent, Inc., Tokyo 150-0042, Japan; otani_mayu@cyberagent.co.jp
[3] Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University, Kyoto 606-8507, Japan; chu@i.kyoto-u.ac.jp
* Correspondence: y-hirota@is.ids.osaka-u.ac.jp
† This paper is an extended version of our paper published in a paper entitled "Visual Question Answering with Textual Representations for Images", which was presented at Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 11–17 October 2021.

Abstract: How far can textual representations go in understanding images? In image understanding, effective representations are essential. Deep visual features from object recognition models currently dominate various tasks, especially Visual Question Answering (VQA). However, these conventional features often struggle to capture image details in ways that match human understanding, and their decision processes lack interpretability. Meanwhile, the recent progress in language models suggests that descriptive text could offer a viable alternative. This paper investigated the use of descriptive text as an alternative to deep visual features in VQA. We propose to process description–question pairs rather than visual features, utilizing a language-only Transformer model. We also explored data augmentation strategies to enhance training set diversity and mitigate statistical bias. Extensive evaluation shows that textual representations using approximately a hundred words can effectively compete with deep visual features on both the VQA 2.0 and VQA-CP v2 datasets. Our qualitative experiments further reveal that these textual representations enable clearer investigation of VQA model decision processes, thereby improving interpretability.

Keywords: visual question answering; textual representations; data augmentation; interpretability; vision-and-language

check for updates

Citation: Hirota, Y.; Garcia, N.; Otani, M.; Chu, C.; Nakashima, Y. A Picture May Be Worth a Hundred Words for Visual Question Answering. *Electronics* **2024**, *13*, 4290. https://doi.org/10.3390/electronics13214290

Academic Editors: Panayiotis Bozanis and Leonidas Akritidis

Received: 3 October 2024
Revised: 24 October 2024
Accepted: 29 October 2024
Published: 31 October 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Understanding the visual world through language has been a fundamental challenge in artificial intelligence. Computer vision systems have made remarkable progress in tasks like object detection [1] and scene understanding [2,3], primarily relying on deep visual features—mathematical representations of images learned through neural networks. However, these features face two critical limitations [4–6]: they often struggle to capture complex semantic relationships in images, and their decision-making process remains opaque to humans [7].

The English proverb "A picture is worth a thousand words" suggests that visual information can convey meaning more efficiently than verbal descriptions. However, recent advances in natural language processing, particularly the emergence of powerful language models, such as Transformer [8] and Transformer-based models [9–12], have demonstrated remarkable capabilities in understanding and reasoning about textual information, suggesting that concise textual descriptions might serve as an effective alternative to image representation.

Current approaches to visual understanding tasks, such as Visual Question Answering (VQA), have explored various ways to represent images. Traditional systems rely on deep

visual features extracted by object recognition models [1], while more recent approaches utilize Transformer-based models [13–20] like Vision Transformer [21] to learn representations from image–text pairs. Despite their widespread adoption, these visual representations often struggle to capture detailed semantic relationships and lack human interpretability, limiting their effectiveness in complex reasoning tasks like VQA.

In this paper, we explored a different approach to VQA that moves away from conventional visual feature-based methods. Instead of focusing on improving image feature extraction or visual processing techniques, we investigated the potential of using pure textual descriptions to represent image content. Specifically, we investigated whether well-crafted textual descriptions can serve as an effective alternative to traditional visual features, potentially offering benefits in both performance and interpretability.

To address these challenges, we propose a novel approach that replaces image–question pairs with image description–question pairs in VQA tasks. Our key research questions are as follows:

- Can textual representations compete with or outperform deep visual features in VQA tasks?
- How many words are actually necessary to effectively represent an image for machine understanding?
- Can data augmentation techniques enhance the performance of text-based image understanding?

We used RoBERTa [9], one of the most well-known Transformer-based language models, as our VQA model. The input description–question pairs were jointly fed into the model to predict an answer. In addition, with the success of data augmentation methods on both VQA and NLP tasks [22–26], we investigated the use of synthetic samples on language-only representations. As the aim of the study was to explore the viability of language-only representations in VQA, we relied on already annotated descriptions from two standard datasets [27,28]. Automatically generating the image descriptions, although a necessary future step, is out of the scope of this paper.

Our work makes three key contributions:

- We demonstrate that textual representations using approximately 100 words can match or exceed the performance of deep visual features on standard VQA benchmarks, challenging conventional wisdom about image representation.
- We introduce a more interpretable approach to VQA, where the system's decision-making process can be readily understood by examining the textual descriptions it uses.
- We present novel data augmentation techniques adapted for text-only VQA, including a particularly effective back translation method for questions that significantly improves performance.

The remainder of this paper is organized as follows: Section 2 reviews related work in VQA, image representation, and data augmentation. Section 3 details our proposed approach and methodology. Section 4 presents our experimental results and analysis. Section 5 provides qualitative analysis and discussion of our findings, Section 6 discusses the limitations of our approach and the benefits from our results, and Section 7 concludes with implications for future research.

2. Related Work

Image Representations for VQA. Image representation plays a crucial role in vision-and-language tasks. Traditional VQA approaches [29–34] rely on deep visual features extracted by object detectors like Faster R-CNN [1], where each feature vector captures information about a specific image region. Recent models utilizing Transformer-based architectures [21] have shown promising results [13–15,35]. However, these approaches face two key challenges: they often learn superficial correlations rather than true vi-

sual understanding [4–6], and their internal representations remain difficult for humans to interpret [7].

On the other hand, there has been a growing trend in studies utilizing textual representations of images [18–20,36–40]. These approaches offer several advantages: they can capture semantic relationships more explicitly, provide human-interpretable representations, and leverage recent advances in language understanding. Many works have adapted Transformer models to fuse visual and textual information [13,16–19,41], achieving high performance through pre-training with image–caption pairs. For example, Wu et al. [38] generates question-relevant captions to provide additional context for answering. While these approaches demonstrate the value of textual information, they typically treat text as supplementary to visual features rather than as the primary representation medium. This leaves open the question of whether textual descriptions alone could serve as effective image representations.

Recent works have shown promising solutions to various VQA challenges: SCLSM [42] effectively addresses language bias through contrastive learning, Atlantis [43] successfully integrates aesthetic features for sentiment analysis, and studies on out-of-distribution detection [44] improve system reliability. While these approaches advance VQA through different learning strategies, our work takes a fundamentally different direction by investigating the potential of pure textual representations for image understanding.

Data Augmentation for VQA. Data augmentation techniques for VQA have primarily focused on addressing language bias [4–6], where models tend to exploit superficial correlations between questions and answers in the training set. Chen et al. [23] demonstrated the effectiveness of counterfactual sample synthesis by manipulating critical parts of input images, while Gokhale et al. [22] showed that systematic manipulation of both images and questions can improve model robustness. While these approaches have proven effective, they require computationally expensive image manipulation.

Our work adapts these counterfactual generation principles to operate directly on textual descriptions, preserving the benefits of data augmentation while reducing computational costs. Furthermore, we extend these ideas by incorporating techniques from NLP-based data augmentation, creating a framework that leverages the strengths of both VQA-specific and text-specific augmentation methods.

Data Augmentation for NLP. The NLP community has developed diverse data augmentation strategies to enhance model performance. Back translation [25,26,45] stands as one of the fundamental techniques where text is transformed through intermediate languages to create semantically equivalent but linguistically diverse samples. Another significant approach is EDA (Easy Data Augmentation) [24], which has demonstrated success in various text classification scenarios [46–50], particularly when training data are limited. Recent methods have also explored contextual augmentation, where words are strategically replaced or inserted based on their semantic context.

Given the text-centric nature of our approach, we investigated how these established NLP augmentation techniques can benefit VQA tasks when applied to our textual representations. This novel integration of NLP-specific augmentation methods into VQA presents an opportunity to leverage well-studied text manipulation strategies in a new context.

Distinction from Previous Work. While previous studies have made significant progress in VQA, our approach differs in several key aspects. Unlike works that use textual information as supplementary input [38,39], we explored the potential of using text as the primary representation of visual content. Moreover, while existing data augmentation methods [22,23] focus on image manipulation, our approach uniquely combines VQA-specific and NLP-based augmentation techniques in a text-only setting. This not only reduces computational costs but also provides better interpretability, as all operations are performed in the human-readable text domain.

3. Approach

We present a text-based model to explore the potential of language-only representations for VQA, as shown in Figure 1. The input, comprising a question and a detailed image description, is processed by a Transformer model with multiple self-attention layers. The Transformer's output is then passed to a classifier to predict the answer. Additionally, we employ data augmentation techniques to expand and diversify the training set.

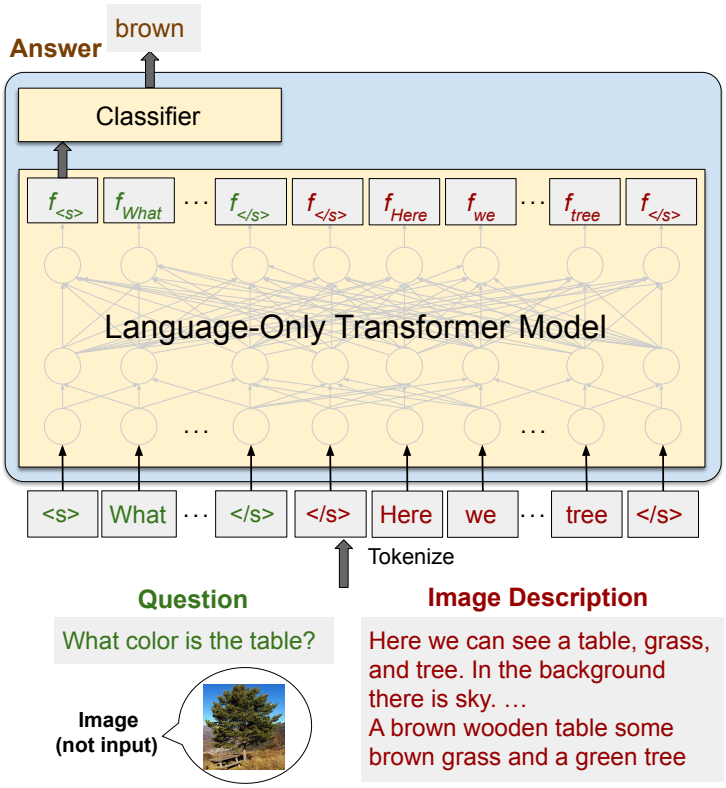


Figure 1. Model overview. Our language-only model takes a question and a description as the input of a language-only Transformer model and predicts an answer accordingly.

3.1. Language-Only Data

Our language-only VQA framework utilizes the following: (1) questions and answers from standard VQA datasets, (2) image descriptions representing the image content, and (3) synthetic data generated through data augmentation techniques.

Questions and Answers. We use the VQA-CP v2 [4] and VQA 2.0 [51] datasets for the questions and answers. VQA 2.0 consists of 1.1 M question–answer pairs across 204 K images, while VQA-CP v2 contains 603 K question–answer pairs from 219 K images. Both datasets use images from the MSCOCO dataset [52]. Although VQA 2.0 is a standard for natural image VQA, it has been shown to exhibit strong statistical biases in its training distribution [4–6], allowing models to achieve high accuracy by focusing on the first few words of a question. VQA-CP v2 mitigates this issue by reorganizing the training and validation splits.

Image Descriptions. We source image descriptions from two corpora: COCO captions [27] and Localized Narratives [28]. COCO captions provide five short captions per image in the MSCOCO dataset [52], with an average of 10.5 words per caption. These captions focus on describing key elements in the scene, omitting less important details. Localized Narratives, on the other hand, offer detailed image descriptions for several datasets, including MSCOCO. Annotators generate these narratives by describing the image aloud while highlighting the relevant regions with a mouse, capturing both prominent and minor objects. This results in more comprehensive descriptions, averaging 42.9 words per narrative.

Synthetic Data. To enhance the diversity of the training set in our language-only framework, we generate synthetic samples using data augmentation techniques, categorized into two types: Data Augmentation for VQA (Section 3.2) and Data Augmentation for Language (Section 3.3). For Data Augmentation for VQA (DAV), inspired by recent multimodal augmentation methods [22,23], which generate new images and questions by altering objects in an image or question, we adapt these methods for our text-only approach by creating synthetic descriptions instead of modifying images. For Data Augmentation for Language (DAL), we employ well-established NLP techniques to enhance language tasks, drawing from methods such as EDA, back translation, and others [24–26,45].

3.2. Data Augmentation for VQA

We utilize data augmentation methods for VQA [22,23] for our language-only input for two reasons: (1) to generate diverse training samples that force the model to focus on essential information rather than superficial patterns, and (2) to achieve this in a computationally efficient way by manipulating text rather than images. While previous approaches [22,23,53,54] required complex image manipulation through masking or GANs [55], our text-based methods achieve similar diversity through simple word-level operations.

Let s denote a training triplet, i.e., $s = (\mathbf{q}, \mathbf{d}, A)$, where $\mathbf{q} = [q_1, \cdots, q_Q]$ is a question sequence with Q tokens, associated with a question type t, and $\mathbf{d} = [d_1, \cdots, d_D]$ is a description sequence with D tokens. The set $A = \{a_1, \cdots, a_N\}$ contains N ground-truth answers, with N being at most 10 for VQA 2.0 and VQA-CP v2 datasets, varying across samples. We propose four data augmentation techniques: (1) hypernym and hyponym replacement, (2) color inversion, (3) adversarial replacement, and (4) counterfactual samples. Examples are illustrated in Figure 2.

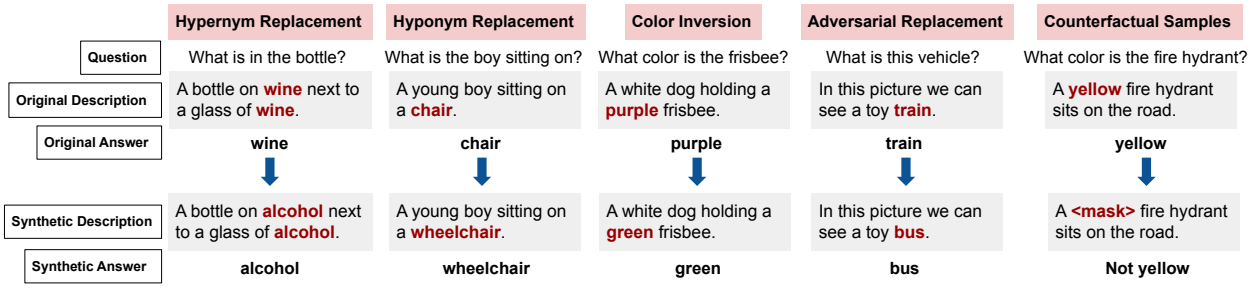


Figure 2. Examples of synthetic samples generated using our proposed data augmentation techniques for VQA.

Hypernym and Hyponym Replacement. Following [22], we use hypernym and hyponym replacement to create semantically meaningful variations in image descriptions while preserving their basic truth value. For example, replacing "fruit" with its hypernym "food" encourages the model to generalize across broader categories, while replacing it with its hyponym "apple" allows the model to handle specific instances. In a real-world scenario, if the description contains the word "car", it might be replaced with "vehicle" (hypernym) to generalize, or "sedan" (hyponym) to specialize. This technique forces the model to learn more abstract or specific concepts, which reduces reliance on narrow language patterns and helps in combating biases that arise from overfitting to common words in the training data.

To generate new samples, let $A_{\mathbf{d}}$ be the set of ground-truth answers a that appear in $\mathbf{d}$. We replace $a \in A_{\mathbf{d}}$ with its hypernym $h_e(a)$ (or hyponym $h_o(a)$) in both A and $\mathbf{d}$. The new triplet for hypernym replacement, s_{h}, is defined as follows:

$$s_{\mathrm{h}} = (\mathbf{q}, \mathbf{d}_{\mathrm{h}}, A_{\mathrm{h}}) \tag{1}$$

$$A_{\mathrm{h}} = A \setminus A_{\mathbf{d}} \cup \{h_e(a)\}_{a \in A_{\mathbf{d}}} \tag{2}$$

$$\mathbf{d}_{\mathrm{h}} = [d_1, \cdots, d_L] \quad \text{where } d_i = h_e(a) \text{ if } d_i = a, \quad \forall a \in A_{\mathbf{d}} \tag{3}$$

Similarly, $h_o(a)$ is used for hyponym replacement. If $h_e(a) \in A$ (or $h_o(a) \in A$), we skip generating s_h to avoid duplicates. Hypernyms and hyponyms are identified using WordNet [56].

Color Inversion. For color inversion [22], we substitute a color word in a description with another color word. For example, if the original description states that "the car is white", we replace "white" with "black", resulting in "the car is black". This method helps the model handle different color schemes and ensures that it does not overly rely on specific color cues, making it more robust in tasks that involve color-based questions or object recognition.

We define a set of K color words, $C = \{c_1, \cdots, c_K\}$, and a set of question types related to colors, T_C (T_C = {"what color", "what color are the", "what color is", "what color is the"}). For a training triplet s with question type t, if $t \in T_C$ and a ground-truth color answer $a \in A \cap C$ appears in the description (i.e., $d_i = a$ for some i), we randomly replace the color word a with a different color $c \neq a$. The new training triplet s_c is as follows:

$$s_\mathrm{c} = (\mathbf{q}, \mathbf{d}_\mathrm{c}, A_\mathrm{c}) \tag{4}$$

$$A_\mathrm{c} = A \setminus \{a\} \cup \{c\} \tag{5}$$

$$\mathbf{d}_\mathrm{c} = [d_1, \cdots, d_L] \text{ with } d_i = c \text{ if } d_i = a \tag{6}$$

Adversarial Replacement. For Yes/No samples, we replace object words in the description with adversarial words. For instance, if the original description states "the cat is on the sofa" and the question is "Is the cat on the sofa?", we might replace "sofa" with "chair", resulting in "the cat is on the chair", forcing the model to change the answer from "yes" to "no". This helps the model learn to better distinguish between subtle changes in object references and makes it more robust in answering questions accurately when similar but different objects are involved.

For Yes/No samples, i.e., $s = (\mathbf{q}, \mathbf{d}, A)$ where $\{\text{yes}, \text{no}\} \cap A \neq \emptyset$, we replace object words $o \in O$ in the description $\mathbf{d}$ with adversarial words. O is the set of 80 object classes in MSCOCO [52]. Following [22], we define an adversarial word, $w_\mathrm{adv}(o)$, as the word most similar to o but with a different meaning. If o (or its synonyms) appear in $\mathbf{q}$, the answer is switched from "yes" to "no"; otherwise, the answer remains unchanged. The new training triplet s_a is as follows:

$$s_\mathrm{a} = (\mathbf{q}, \mathbf{d}_\mathrm{a}, A_\mathrm{a}) \tag{7}$$

$$A_\mathrm{a} = \begin{cases} \{\text{no}\} & \text{if } o \text{ is in } \mathbf{q} \\ A & \text{otherwise} \end{cases} \tag{8}$$

$$\mathbf{d}_\mathrm{a} = [d_1, \cdots, d_L] \text{ with } d_i = w_\mathrm{adv}(o) \text{ if } d_i = o, \tag{9}$$

where $w_\mathrm{adv}(o)$ is selected as the closest word to $o \in O$ based on the Euclidean distance between their GloVe embeddings [57].

Unlike [22], we avoid generating adversarial samples from questions, as altering a word in the question results in a new answer that cannot be automatically inferred, e.g., "How many bins?" →"How many pens?".

Counterfactual Samples. Counterfactual samples [23] are generated by modifying parts of the question or description to create alternative scenarios. For example, if the original question is "What color is the car?", and the description states "The car is red", we modify the description to "The car is blue" to generate a counterfactual sample. This forces the model to understand changes in the input and adjust its predictions accordingly, improving its ability to handle ambiguous or altered situations.

Specifically, we generate counterfactual training samples (CSS) by adapting the method in [23] for language-only description–question pairs. For a given training triplet $s = (\mathbf{q}, \mathbf{d}, A)$, we create counterfactual samples s_{cssq} and $s\text{css}_d$ from the query and description.

$$s_{\text{css}_\text{q}} = (\mathbf{q}_{\text{css}_\text{q}}, \mathbf{d}, A_{\text{css}_\text{q}}) \tag{10}$$

$$s_{\text{css}_\text{d}} = (\mathbf{q}, \mathbf{d}_{\text{css}_\text{d}}, A_{\text{css}_\text{d}}) \tag{11}$$

To generate s_{cssq}, we input $\mathbf{q}$ and $\mathbf{d}$ into a trained Transformer VQA model $\mathcal{M}$ and use Grad-CAM [58] to find the contribution of each word in $\mathbf{q}$ to the answer set A. The top-D words with the highest contribution form the critical set Ωq. Two new questions, $\mathbf{q}^{+}_{\text{css}_\text{q}}$ and $\mathbf{q}^{-}_{\text{css}_\text{q}}$, are created by masking words not in Ω_q and masking the words in Ω_q, respectively.

$$\begin{aligned} \mathbf{q}^{+}_{\text{css}_\text{q}} &= [q_1, \cdots, q_L] \\ &\quad \text{with } q_i = \texttt{<mask>} \quad \text{for all } q_i \notin \Omega_\text{q} \end{aligned} \tag{12}$$

$$\begin{aligned} \mathbf{q}^{-}_{\text{css}_\text{q}} &= [q_1, \cdots, q_L] \\ &\quad \text{with } q_i = \texttt{<mask>} \quad \text{for all } q_i \in \Omega_\text{q}, \end{aligned} \tag{13}$$

where $\mathbf{q}^{-}_{\text{css}_\text{q}}$ serves as the question $\mathbf{q}_{\text{css}_\text{q}}$ for the CSS sample; the first few words that indicate the question type (e.g., "what color is") are not masked, as in [23].

To determine A_{css_q}, we feed $\mathbf{q}^{+}_{\text{css}_\text{q}}$ and $\mathbf{d}$ back into $\mathcal{M}$ to score each candidate answer. The top-J scoring answers are excluded from the original answer set. Specifically, letting $\mathcal{M}_J(\mathbf{q}^{+}_{\text{css}_\text{q}}, \mathbf{d})$ represent the top-J answers, the new ground-truth answers are given by

$$A_{\text{css}_\text{q}} = A \setminus \mathcal{M}_J(\mathbf{q}^{+}_{\text{css}_\text{q}}, \mathbf{d}). \tag{14}$$

For s_{css_d}, the values of Ω_d, $\mathbf{d}^{+}_{\text{css}_\text{d}}$, $\mathbf{d}^{-}_{\text{css}_\text{d}}$, and $M_N(\mathbf{q}, \mathbf{d}^{+}_{\text{css}_\text{d}})$ are determined using the same process.

3.3. Data Augmentation for Language

Given our text-only approach, we leverage established NLP augmentation techniques that have proven effective in maintaining semantic meaning while introducing linguistic diversity. We specifically select three complementary techniques: EDA for simple yet effective transformations, back translation for generating naturally varied paraphrases, and word replacement/insertion via contextual word embedding for ensuring semantic coherence. These methods work together to create a robust training set that helps the model better understand the relationship between textual descriptions and visual concepts. Each technique is applied to either the description or the question in the input triplet s to create new samples:

$$s_{\text{nlp}_\text{q}} = (\mathbf{q}_{\text{nlp}}, \mathbf{d}, A) \tag{15}$$

$$s_{\text{nlp}_\text{d}} = (\mathbf{q}, \mathbf{d}_{\text{nlp}}, A) \tag{16}$$

where $\mathbf{q}_{\text{nlp}}$ and $\mathbf{d}_{\text{nlp}}$ represent the question and description after applying one of the transformations below. Examples of generated synthetic samples are provided in Table 1.

EDA (Easy Data Augmentation) [24] includes the following four operations:

- Synonym Replacement randomly selects n words from the sentence and substitutes them with their synonyms.
- Random Insertion inserts a random synonym of a randomly chosen word into a random position in the sentence.
- Random Swap selects two words from the sentence and swaps their positions.
- Random Deletion removes words from the sentence with a probability of p.

Table 1. Examples of Data Augmentation for Language when applied to questions.

Original	**Is This an Ocean Area?**	**What Is the Giraffe Standing Behind?**
EDA (Synonym Replacement)	Is this an ocean region?	What is the camelopard standing behind?
EDA (Random Insertion)	Is this sea an ocean area?	What is the giraffe abide standing behind?
EDA (Random Swap)	Is this ocean an area?	What is the standing giraffe behind?
EDA (Random Deletion)	Is this an area?	What is the standing behind?
Back Translation	Is it a maritime area?	What's behind the giraffe?
Contextual Word Replacement	Is this an ocean top?	What is the giraffe tree behind?
Contextual Word Insertion	Is this an urban ocean area?	What is the giraffe standing silently behind?

For example, using synonym replacement, the sentence "The cat is sitting on the sofa" can become "The feline is sitting on the couch". This introduces variations in the input data while keeping the meaning intact, making the model more robust to paraphrases and improving its generalization across different language expressions. For each sample, one of these four operations is applied at random.

Back Translation [25,26] translates a sentence into another language and then translates it back into the original language. For instance, translating "The dog is playing in the garden" to German and back into English might produce "The dog plays in the yard." This method introduces linguistic variety while preserving meaning, which helps to reduce bias introduced by specific language patterns in the training data. By generating paraphrases that use different sentence structures and vocabulary, the model is exposed to a broader range of linguistic expressions, reducing overfitting to common phrases and thus mitigating language bias. For back translation, we implement it using the Python library nlpaug [59], which translates a sentence into German and back into English. If the sentence remains unchanged after translation, we discard it.

Contextual Word Replacement/Insertion replaces or inserts words based on the surrounding context. For example, if the original sentence is The dog is running in the park, the word "park" might be replaced with "field" or "garden" based on the context. This helps the model to generalize its understanding of different environments while maintaining the overall meaning of the sentence, making it more robust to variations in language and context. To obtain contextually appropriate words for replacement or insertion, we leverage the Python nlpaug library [59], selecting random words from the description or question and replace or insert the most contextually similar words. Pre-trained XLNet [60] is used for generating the contextual word embeddings needed for these transformations.

3.4. Language-Only VQA Model

Unlike most VQA models that take image–question pairs as input, our language-only VQA model uses description–question pairs. For the original VQA triplet $(\mathbf{q}, I, A)$, where I is the image related to the question, we generate the image description $\mathbf{d}$ by combining the narrative from Localized Narratives with the captions from COCO captions. Following the sequence format of the RoBERTa's implementation in Huggingface [61], the question and image description are then merged into a single sequence, $\mathbf{l}$, by inserting a classifier token <s> at the start, and sentence-ending tokens </s> as follows:

$$\mathbf{l} = \texttt{<s>} + \mathbf{q} + \texttt{</s>} + \texttt{</s>} + \mathbf{d} + \texttt{</s>} \tag{17}$$

where "+" denotes concatenation. The resulting input sequence is then fed into our Transformer-based language model $\mathcal{T}$, producing the sequence of embeddings $\mathbf{f}$, i.e.,

$$\mathbf{f} = \mathcal{T}(\mathbf{l}). \tag{18}$$

Then, the embedding corresponding to the classifier token <s> is passed into the classifier $\mathcal{C}$ to generate the final prediction:

$$\rho = \mathcal{C}(\mathbf{f}_{<s>}) \tag{19}$$

List of Abbreviations

We summarize the abbreviations used in our approach and its explanations.

- CSS: Counterfactual Samples Synthesizing;
- EDA: Easy Data Augmentation;
- VQA: Visual Question Answering;
- NLP: Natural Language Processing.

Summary: This section introduced our approach of using textual representations and a Transformer-based model, with the support of data augmentation techniques to enhance performance.

4. Experiments

We conducted five main experiments to evaluate the effectiveness of textual representations in VQA: (Section 4.1) comparing various image descriptions, (Section 4.2) assessing different data augmentation techniques, (Section 4.3) comparing textual representations with deep visual features, (Section 4.4) examining the effect of data augmentation on questions for models using deep visual features, and (Section 4.5) comparing different language-only Transformers.

Setup. We chose RoBERTa [9] as the primary Transformer model for our experiments due to its strong performance on a wide range of natural language processing tasks, including question answering and its improvements over BERT [10], such as better pre-training techniques and removal of the next sentence prediction objective. These optimizations lead to better generalization and higher accuracy, making RoBERTa well-suited for VQA tasks where both questions and image descriptions must be processed efficiently. While other models like BERT and XLNet were also viable, RoBERTa consistently showed superior results in our experiments (Section 4.5) . For RoBERTa hyperparameters, we used the large variant with 24 layers and 355 million parameters. The model was trained with a batch size of 32 and a learning rate of 1×10^{-5}, using the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The maximum sequence length for inputs was set to 512 tokens, and we used dropout with a rate of 0.1 for regularization. Training was performed for 10 epochs, with early stopping based on validation loss. As a classifier, we used a multi-layer perceptron with two fully-connected layers and the Swish activation function [62] between them. We used softmax cross entropy over the answer vocabulary for the loss function. The parameters for the data augmentation techniques were as follows: for counterfactual samples, we set $D = 10$ to select critical words and $J = 5$ to assign new answers. For EDA, the rate of words to be changed was set to 0.1, as recommended in the original paper [24]. Unless otherwise specified, the input to our model included the entire sequence of the question, narrative, and five captions. Results were presented in terms of accuracy, as in [51].

4.1. Describing an Image with Words

This experiment explored how different types and lengths of image descriptions impact VQA performance. The goal was to determine whether more detailed descriptions improve accuracy and to assess how much information is needed for optimal results. Understanding this helps to clarify the relationship between the richness of image descriptions and the model's ability to answer questions effectively.

Image descriptions. We began by evaluating the performance of different language-only inputs, including the following: only the question, the question with one to five randomly selected captions from COCO captions, the question with a narrative, and the entire input (question, narrative, and five captions). The results, reported in Table 2, also include the average sequence length.

The complete input, which combined the narrative with five captions, averaging 95.3 tokens per sample, achieved the highest performance, indicating that both datasets provided complementary information useful for VQA. When comparing captions with narratives, captions resulted in better accuracy with fewer words. For example, using just two captions surpassed the performance of narratives, even though the word count was nearly half. This suggests that the VQA dataset contains a large number of questions about general image content rather than specific details, as COCO captions tend to focus on prominent elements of the scene, unlike narratives. In other words, most questions that people ask about an image pertain to its key elements. This tendency for humans to emphasize the prominent parts while overlooking minor details is known as reporting bias [63,64].

Input length. We analyzed the relationship between input length and model accuracy by progressively truncating the image descriptions at test time. The model was trained with the full input (question, narrative, and five captions). To preserve context, we randomly shuffled the sentences before truncating them and then returned them to their original order. The question portion of the input remained intact throughout. The results, shown in Figure 3, indicate a steady decline in accuracy as more words were removed from the input.

Accuracy decreased consistently up to a 60% truncation rate, after which performance dropped more sharply. This suggests that while longer descriptions provided useful context, the model struggled to process the input effectively if too much information was lost. This trend aligns with the concept of diminishing returns, where adding more descriptive information initially improves performance but eventually introduces noise or redundancy. Since COCO captions focus on key elements in the image, truncating beyond a certain point likely removes critical content, leading to a steep decline in accuracy after 60%.

Table 2. Performance of different language-only inputs on the VQA-CP v2 test set. Length represents the mean number of tokens in the image descriptions.

Image Description	Length	Accuracy
None (Question-Only)	-	21.39
One Caption	10.5	35.31
Two Captions	21.0	38.49
Three Captions	31.5	40.09
Four Captions	42.0	41.93
Five Captions	52.5	42.34
Narrative	42.9	36.45
Whole (Narrative + Five Captions)	95.3	43.64

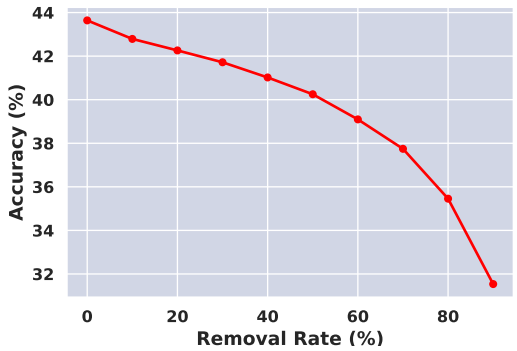


Figure 3. Impact of input sequence truncation on VQA-CP v2 test set accuracy. The graph shows how progressively removing words from the input sequence affected model performance.

4.2. Use of Synthetic Samples

This experiment aimed to evaluate the effectiveness of different data augmentation techniques in improving VQA accuracy. By analyzing how synthetic samples generated through various methods affected the model's performance, we sought to determine which techniques enhanced the model's ability to generalize and handle diverse question types.

We specifically tested the impact of these augmentations on the VQA-CP v2 training set, using the full description with the narrative and five captions as input.

The results for each of the proposed data augmentation techniques are shown in Table 3. While most techniques improved accuracy, back translation applied to questions yielded the largest improvement, particularly for Yes/No questions, with a 17.30 point gain. This boost of 17.30 points was calculated as the difference between the baseline Yes/No accuracy of 45.13% and the Yes/No accuracy of 62.43% achieved when the back translation was applied to the questions. This suggests that paraphrasing through back translation introduced a diversity of linguistic structures while preserving the semantic meaning, helping the model generalize better to question variations. In contrast, methods like hypernym/hyponym replacement or color inversion generated smaller improvements, likely because they involved limited word-level changes that did not sufficiently alter the structure of the descriptions or questions to increase generalization significantly. Contextual word replacement for descriptions even led to a performance drop, likely due to the insertion of words that were semantically less relevant, confusing the model during training.

A notable observation is that applying data augmentation to questions consistently produced better results than applying it to descriptions. Among the DAV techniques, hyponym replacement stood out as the most effective, with a significant accuracy boost of 1.62. Interestingly, combining hyponym and hypernym replacements did not improve performance beyond hyponym replacement alone. This indicates that, in some instances, combining synthetic samples from different augmentation techniques can be counterproductive, possibly confusing the model.

Table 3. Results of data augmentation techniques on the VQA-CP v2 test set, showing the impact of different data augmentation methods on model accuracy for Yes/No, Number, and Other question types. D indicates techniques applied to image descriptions, while Q indicates application to questions. The Gap column highlights the improvement in accuracy compared to the baseline, where no synthetic data were used.

	Input Data	Num. Synthetic	Num. Total	Yes/No	Number	Other	Overall	Gap
	Narrative + Five Captions	-	438,183	45.13	20.06	49.33	43.64	-
DAV	w/ Hyponym Replacement	132,570	570,753	45.65	25.36	50.52	45.26	+1.62
DAV	w/ Hypernym Replacement	23,869	462,052	47.28	17.69	49.10	43.70	+0.06
DAV	w/ Hyponym and Hypernym Replacement	183,944	622,177	45.80	21.46	51.15	45.06	+1.42
DAV	w/ Color Inversion	19,308	457,491	45.61	19.93	50.60	44.47	+1.06
DAV	w/ Adversarial Word Replacement	169,929	608,112	44.71	19.84	50.03	43.93	+0.29
DAV	w/ Counterfactual Samples	438,183	876,366	44.20	19.84	52.07	44.86	+1.22
DAL	w/ EDA (D)	438,183	876,366	44.68	20.64	50.08	44.02	+0.38
DAL	w/ EDA (Q)	438,183	876,366	46.86	23.50	50.62	45.39	+1.75
DAL	w/ Contextual Word Replacement (D)	438,183	876,366	44.69	19.40	48.91	43.18	−0.46
DAL	w/ Contextual Word Replacement (Q)	438,183	876,366	46.09	22.49	49.10	44.16	+0.52
DAL	w/ Contextual Word Insertion (D)	438,183	876,366	45.15	19.31	48.86	43.27	−0.37
DAL	w/ Contextual Word Insertion (Q)	438,183	876,366	45.86	21.44	51.10	45.05	+1.41
DAL	w/ Back Translation (D)	438,183	876,366	45.28	21.01	50.89	44.70	+1.06
DAL	w/ Back Translation (Q)	293,811	731,994	62.43	27.15	51.84	51.16	+7.52

4.3. Comparison Against Deep Visual Features

This experiment evaluated the effectiveness of text-based representations in comparison to models using deep visual features. The key question was whether language-only representations can encapsulate the necessary visual information as effectively as traditional visual features. The objective was to assess the potential of text-based approaches as a viable alternative or complement to deep visual features in tasks like VQA, where the understanding of both visual and textual elements is crucial.

We compared our language-only representations (question, narrative, and five captions) with top VQA models that use deep visual features on the VQA-CP v2 and VQA 2.0

datasets. To ensure fairness, we excluded models designed to address language bias [32–34], as these techniques can be applied to any model, including ours. Additionally, we did not use data augmentation for this comparison. Despite the detailed visual information captured by models like VisualBERT, which is trained on image–text pairs, the results in Table 4 demonstrate that language-based models perform competitively. Our approach surpassed many deep visual feature baselines, indicating that well-annotated textual descriptions can effectively capture essential visual details.

One possible explanation for the strong performance of our model is that human-generated image descriptions tend to focus on the most relevant aspects, filtering out unnecessary details that deep visual features might capture. Moreover, text-based representations provide a level of interpretability that may help align the question and description more accurately. However, NSM [65] achieves slightly better results on VQA-CP v2, likely due to its ability to capture more intricate visual details, which can be challenging to express purely through text. Overall, this comparison highlights the advantages of textual representations in tasks where high-level semantic understanding is crucial. However, deep visual features may still hold an edge in tasks requiring precise spatial reasoning or detailed object localization, areas where textual descriptions may fall short.

Table 4. Comparison of language-only models with deep visual feature-based models, showing the performance of various models on the VQA-CP v2 test and VQA 2.0 validation sets. Results marked with * are our re-implementations.

	VQA-CP v2 Test				VQA 2.0 Val			
Model	**Yes/No**	**Number**	**Other**	**Overall**	**Yes/No**	**Number**	**Other**	**Overall**
HAN [66]	52.25	13.79	20.33	28.65	-	-	-	-
MuRel [32]	42.85	13.17	45.04	39.54	-	-	-	65.14
UpDn [29]	42.27	11.93	46.05	39.74	81.18	42.14	55.66	63.48
ReGAT [67]	-	-	-	40.42	-	-	-	67.18
BAN * [30]	43.14	13.63	46.92	40.74	83.19	48.13	57.52	65.93
VisualBERT * [20]	43.30	15.07	47.83	41.51	84.55	48.19	57.29	66.33
NSM [65]	-	-	-	45.80	-	-	-	-
Ours (Narrative + Five Captions)	45.13	20.06	49.33	43.64	87.91	56.47	59.43	69.74

4.4. Back Translation for Other Models

The results in Section 4.2 demonstrate that back translation significantly boosted accuracy in our language-only setting. To assess whether this technique can generalize to other models, we applied back translation to standard VQA models, such as BAN [30] and VisualBERT [20], by adding synthetic back-translated samples to the training set. This experiment aimed to test the generalizability of back translation as a data augmentation technique. We investigated whether the improvements seen in our language-only model also benefited models that incorporate deep visual features, helping to establish back translation as an effective tool across diverse VQA architectures.

We present the results in Table 5. Training BAN and VisualBERT with synthesized back-translated samples significantly boosted performance, with improvements of 2.83 and 5.06 points, respectively. This technique benefited all question types, particularly Yes/No questions, where the improvement was around 12.65 points. These findings align with our model's results, demonstrating that adding synthetically varied questions with the same meaning is highly effective for enhancing model performance.

Table 5. Results of applying back translation to different VQA models in the VQA-CP v2 test set. Gap denotes the improvement achieved by training with synthetic back-translated samples.

Model	Yes/No	Number	Other	Overall	Gap
BAN [30]	43.14	13.63	46.92	40.74	-
w/BT	47.87	16.27	48.76	43.57	+2.83
VisualBERT [20]	43.30	15.07	47.83	41.51	-
w/BT	55.95	17.11	49.74	46.57	+5.06

4.5. Comparison of Language Transformers

In this experiment, we compared the performance of various language-only Transformer models, including BERT [10], XLNet [60], and RoBERTa [9], in both their base and large versions. All models were evaluated using the same input—question, narrative, and five captions. The goal of this experiment was to determine which Transformer architecture is most effective for handling textual representations in VQA. By analyzing the performance of different models, we aimed to gain insights into how various language models manage the complexity of this task.

The results are presented in Table 6. All models exhibited similar performance patterns, with XLNET large achieving the highest accuracy, outperforming RoBERTa large by 0.59%. However, the computational time for XLNET large was approximately 2.7 times longer than for RoBERTa large. Given the minimal accuracy difference, RoBERTa large was a more efficient choice, offering comparable results while significantly reducing training time.

Table 6. Performance comparison of language-only Transformer models on the VQA-CP v2 test set. **Bold** values indicate the highest performance.

Model	Yes/No	Number	Other	Overall
BERT base	42.78	17.53	46.99	41.27
BERT large	42.72	17.43	48.47	42.06
XLNET base	43.49	17.61	48.45	42.30
XLNET large	44.58	**20.67**	**50.52**	**44.23**
RoBERTa base	44.39	17.46	48.74	42.70
RoBERTa large	**45.13**	20.06	49.33	43.64

Summary: Our experiments demonstrate that language-only models, when paired with rich textual descriptions and effective data augmentation, can compete with deep visual feature-based models in VQA tasks. Detailed captions enhance performance, while back translation significantly improves accuracy, especially for Yes/No questions. We also found that models like RoBERTa large provide a strong balance between computational efficiency and accuracy, making them ideal for language-only VQA tasks. Overall, these findings highlight the potential of textual representations as a robust alternative to deep visual features, with wide applicability across different VQA models.

5. In-Depth Analysis of Textual and Visual Representations in VQA

In this section, we provide an in-depth analysis of the strengths and characteristics of textual representations for images, compared to deep visual features. Specifically, we focus on two key analyses: (Section 5.1) examining the overlap in predictions between our language-only model and models using deep visual features, and (Section 5.2) conducting a qualitative analysis of visual examples. For both analyses, our model's input is a combination of the narrative and five captions.

5.1. Error Analysis

We compared our text-only model with models that incorporate deep visual features, examining whether they make similar or different mistakes. We use BAN [30] and VisualBERT [20] as representative models for deep visual features. BAN is a strong pre-Transformer

model, while VisualBERT uses multimodal Transformers and integrates both image and caption inputs, making it an intermediate between BAN and our text-only approach.

Figure 4 shows the consistency of correct and incorrect answers across the models. We used a binary accuracy metric for this comparison, where a prediction is considered correct if it matches any ground-truth answer, as opposed to the original non-binary metric used in the main paper [51]. Our model demonstrated high consistency, with similar rates of correct/incorrect predictions for both BAN and VisualBERT. For 80% of the questions, all models gave consistent results (blue part). For 12% of the questions, only our model answered correctly (orange), while this dropped to 8% where only the baseline models answered correctly (red). This supports the idea that textual image representations can compete with deep visual features.

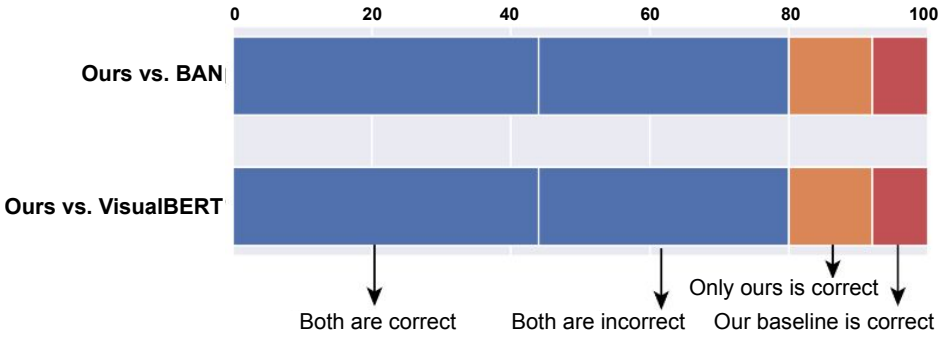


Figure 4. Answer overlap: comparison of prediction agreement between our language-only model, BAN, and VisualBERT. Bar graphs show proportions of identical or differing answers.

5.2. Qualitative Analysis

We analyzed qualitative examples (Figure 5) comparing the predictions of our text-only model, BAN [30], and VisualBERT [20], to identify cases where our method underperformed and suggest potential improvements.

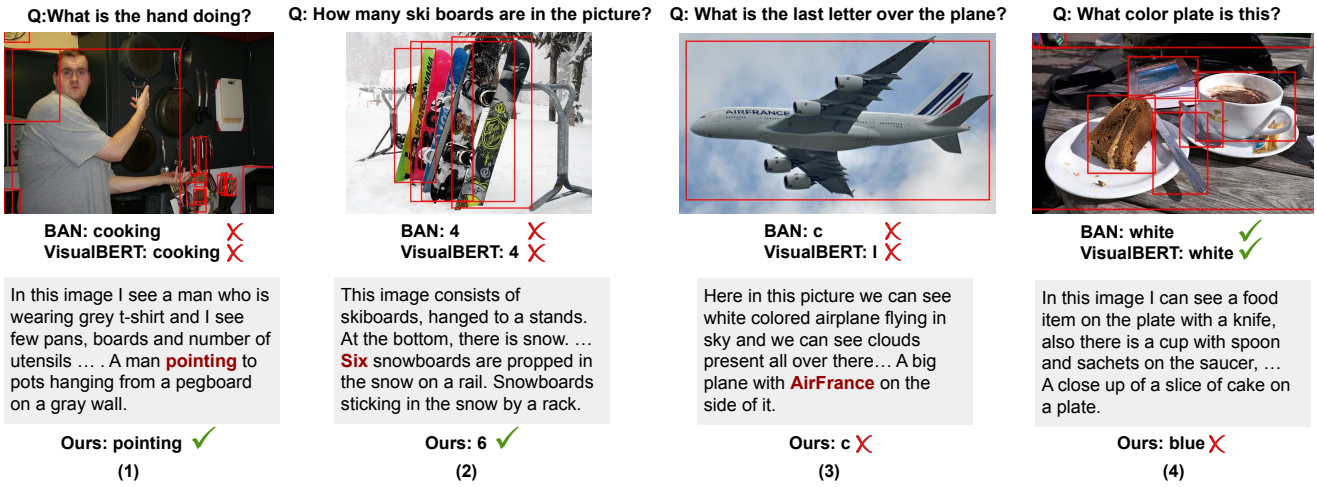


Figure 5. Qualitative comparison: red boxes indicate object detection results by Faster R-CNN with confidence scores over 0.5. Highlighted words in the descriptions correspond to key details relevant to the answers.

In Example (1), our text-only model correctly predicted the man's action based on the description *"A man pointing to pots..."*, while both BAN and VisualBERT incorrectly answered *"cooking"*, likely due to the presence of cooking tools in the scene. This demonstrates how textual representations can more accurately capture actions by relying on explicit information, while deep visual models may misinterpret object presence, underscoring our model's advantage in avoiding irrelevant visual cues.

Example (2) highlights another strength of our method: object counting. The detailed description *"Six snowboards"* enabled our model to correctly predict the answer, whereas the object detectors used in BAN and VisualBERT missed detecting all the relevant objects. This underscores the robustness of our approach in handling precise textual descriptions, which can outperform deep visual features when object detection is incomplete.

However, Example (3) reveals a limitation: despite the description containing the word *"AirFrance"*, our model failed to predict the correct answer. This indicates that our model struggles with precise visual-text alignment, where deep visual features might provide stronger spatial or contextual cues. Improvements could focus on better integrating context from the image to reinforce textual grounding.

Lastly, Example (4) exposes another limitation of our model—handling incomplete or insufficient text input. The lack of essential details in the description led to incorrect predictions. This limitation suggests that future work should aim to enhance the model's ability to infer or handle ambiguity when text descriptions lack critical information.

Overall, this analysis highlights the advantages of our text-based model in capturing detailed, action-related, and count-specific information, often missed by deep visual models. However, it also reveals areas for improvement, particularly in integrating contextual cues from images and addressing gaps in text descriptions for more robust performance.

Summary: In this section, we analyzed the performance of textual representations for VQA compared to models that use deep visual features. Through error analysis, we found that our text-only model can make competitive predictions, often outperforming models that rely on visual features. Qualitative analysis further revealed that textual input excels in capturing actions and counting objects, but struggles with precise visual-text alignment and when descriptions are incomplete. This highlights both the strengths and limitations of text-based models and points to areas for potential improvement.

6. Discussion

Summary of Results and Time Complexity. Our experiments demonstrate that language-only models, particularly when combined with data augmentation techniques like back translation, can achieve competitive performance compared to deep visual models. For instance, back translation improved Yes/No question accuracy by 17.30 points, outperforming most augmentation methods. While deep visual models like NSM slightly outperform our model in some areas, the interpretability and semantic richness of text-based models present a distinct advantage, especially in scenarios requiring human-like reasoning. Additionally, visual feature-based models like BAN and VisualBERT require more time to train than our language-only model. This is due to the additional feature merging module needed to combine image and textual features in these models. For instance, our language-only model trains approximately 3.2 times faster than BAN, making it a more efficient choice in terms of computational time while achieving competitive accuracy.

Comparison between text representation and deep visual features. Directly comparing our language-only model with deep visual feature-based models is challenging due to different inputs—our model uses human-annotated descriptions, while visual models rely on extracted visual features. Human descriptions often align better with the questions, offering a distinct advantage. Despite these differences, our results highlight the parallels and distinctions between text-based and visual features. A key strength of our approach is its interpretability, as human-readable descriptions help explain the model's decisions, making it a valuable baseline for VQA and showing the potential of textual representations as a complement or alternative to visual features, especially when interpretability is key.

Future Research Directions. This study opens up new research avenues for VQA and image understanding. One promising direction is to automatically generate image descriptions as representations, either to supplement or replace deep visual features. This could bridge the gap between textual and visual models, combining the interpretability of text with the precision of deep visual features. Additionally, exploring how textual and visual features can be harmonized to improve models' ability to handle complex reasoning

tasks, such as spatial relationships or fine-grained object recognition, is a logical next step. Another direction could involve developing models that dynamically decide when to rely on text versus visual input, optimizing performance based on the complexity of the task at hand.

Additional baselines. While we acknowledge that including a wider range of baseline models would provide a more comprehensive benchmark, the scope of this study focused on representative models that are well-established in the VQA domain. Future work will aim to incorporate a broader selection of models to further evaluate the generalizability of our approach. Given the model-agnostic nature of our method, we anticipate that the benefits observed in this study will extend to other VQA models as well, providing additional insights into the effectiveness of language-based representations.

7. Conclusions

This paper investigated the use of textual representations of images as an alternative to deep visual features for VQA. We also applied data augmentation techniques to both descriptions and questions to expand the training data and improve diversity. Our experiments showed that the language-only model performs competitively with models using deep visual features. Notably, back translation for questions significantly boosted performance. Our findings suggest that machines do not need overly detailed descriptions to understand images—concise, relevant text is sufficient.

Author Contributions: Conceptualization, Y.H., N.G., M.O., C.C. and Y.N.; methodology, Y.H. and N.G.; software, Y.H.; validation, Y.H. and N.G.; formal analysis, Y.H.; investigation, Y.H.; resources, Y.N.; data curation, Y.H.; writing—original draft preparation, Y.H.; writing—review and editing, N.G., M.O., C.C. and Y.N.; visualization, Y.H.; supervision, Y.N.; project administration, Y.N.; funding acquisition, Y.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by SPS KAKENHI Number JP18H03264 and JP20K19822, and JST CREST Grant Number JPMJCR20D3, Japan.

Data Availability Statement: The data used in this study are publicly available from the following sources: [4,27,51,52].

Conflicts of Interest: Author Mayu Otani is employed by CyberAgent, Inc. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest. The funding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

References

1. Ren, S.; He, K.; Girshick, R.B.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In Proceedings of the 2015 Advances in Neural Information Processing Systems (NeurIPS), 2015, Ontreal, BC, Canada, 7–12 December 2015; pp. 91–99.
2. You, Q.; Jin, H.; Wang, Z.; Fang, C.; Luo, J. Image Captioning with Semantic Attention. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
3. Yao, T.; Pan, Y.; Li, Y.; Qiu, Z.; Mei, T. Boosting Image Captioning with Attributes. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.
4. Agrawal, A.; Batra, D.; Parikh, D.; Kembhavi, A. Don't Just Assume; Look and Answer: Overcoming Priors for Visual Question Answering. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018.
5. Agrawal, A.; Batra, D.; Parikh, D. Analyzing the Behavior of Visual Question Answering Models. In Proceedings of the EMNLP, Austin, TX, USA, 1–5 November 2016.
6. Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; Parikh, D. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
7. Lage, I.; Doshi-Velez, F. Human-in-the-loop learning of interpretable and intuitive representations. In Proceedings of the ICML Workshop on Human Interpretability in Machine Learning, Vienna, Austria, 17 July 2020; Volume 17.

8. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is All you Need. In Proceedings of the 2017 Advances in Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 4–9 December 2017.
9. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* **2019**, arXiv:1907.11692.
10. Devlin, J.; Chang, M.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL-HLT* **2019**, *1*, 2.
11. Taori, R.; Gulrajani, I.; Zhang, T.; Dubois, Y.; Li, X.; Guestrin, C.; Liang, P.; Hashimoto, T.B. Stanford Alpaca: An Instruction-Following LLaMA Model. 2023. Available online: https://github.com/tatsu-lab/stanford_alpaca (accessed on 28 October 2024).
12. Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. Llama: Open and efficient foundation language models. *arXiv* **2023**, arXiv:2302.13971.
13. 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 ICML, PMLR, Baltimore, MD, USA, 17–23 July 2022.
14. Chen, X.; Wang, X.; Changpinyo, S.; Piergiovanni, A.; Padlewski, P.; Salz, D.; Goodman, S.; Grycner, A.; Mustafa, B.; Beyer, L.; et al. Pali: A jointly-scaled multilingual language-image model. In Proceedings of the ICLR, Kigali, Rwanda, 1–5 May 2023.
15. Wang, W.; Bao, H.; Dong, L.; Bjorck, J.; Peng, Z.; Liu, Q.; Aggarwal, K.; Mohammed, O.K.; Singhal, S.; Som, S.; et al. Image as a foreign language: BEiT pretraining for vision and vision-language tasks. In Proceedings of the 2023 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–24 June 2023.
16. Lu, J.; Batra, D.; Parikh, D.; Lee, S. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In Proceedings of the 2019 Advances in Neural Information Processing Systems (NeurIPS), Vancouver, BC, Canada, 8–14 December 2019.
17. Tan, H.; Bansal, M. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In Proceedings of the EMNLP/IJCNLP, Hong Kong, China, 3–7 November 2019.
18. Chen, Y.; Li, L.; Yu, L.; Kholy, A.E.; Ahmed, F.; Gan, Z.; Cheng, Y.; Liu, J. UNITER: UNiversal Image-TExt Representation Learning. In Proceedings of the ECCV, Glasgow, UK, 23–28 August 2020.
19. Su, W.; Zhu, X.; Cao, Y.; Li, B.; Lu, L.; Wei, F.; Dai, J. VL-BERT: Pre-training of Generic Visual-Linguistic Representations. In Proceedings of the ICLR, Addis Ababa, Ethiopia, 30 April 2020.
20. Li, L.H.; Yatskar, M.; Yin, D.; Hsieh, C.J.; Chang, K.W. VisualBERT: A Simple and Performant Baseline for Vision and Language. *arXiv* **2019**, arXiv:1908.03557.
21. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. In Proceedings of the ICLR, Vienna, Austria, 4 May 2021.
22. Gokhale, T.; Banerjee, P.; Baral, C.; Yang, Y. MUTANT: A Training Paradigm for Out-of-Distribution Generalization in Visual Question Answering. In Proceedings of the EMNLP, Online, 16–20 November 2020.
23. Chen, L.; Yan, X.; Xiao, J.; Zhang, H.; Pu, S.; Zhuang, Y. Counterfactual Samples Synthesizing for Robust Visual Question Answering. In Proceedings of the CVPR, Seattle, WA, USA, 14–19 June 2020.
24. Wei, J.W.; Zou, K. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. In Proceedings of the EMNLP/IJCNLP, Hong Kong, China, 3–7 November 2019.
25. Sennrich, R.; Haddow, B.; Birch, A. Improving Neural Machine Translation Models with Monolingual Data. In Proceedings of the ACL, Berlin, Germany, 7–12 August 2016.
26. Yu, A.W.; Dohan, D.; Luong, M.; Zhao, R.; Chen, K.; Norouzi, M.; Le, Q.V. QANet: Combining Local Convolution with Global Self-Attention for Reading Comprehension. In Proceedings of the ICLR, Vancouver, BC, Canada, 30 April–3 May 2018.
27. Chen, X.; Fang, H.; Lin, T.; Vedantam, R.; Gupta, S.; Dollár, P.; Zitnick, C.L. Microsoft COCO Captions: Data Collection and Evaluation Server. *arXiv* **2015**, arXiv:1504.00325.
28. Pont-Tuset, J.; Uijlings, J.R.R.; Changpinyo, S.; Soricut, R.; Ferrari, V. Connecting Vision and Language with Localized Narratives. In Proceedings of the ECCV, Glasgow, UK, 23–28 August 2020.
29. Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; Zhang, L. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In Proceedings of the CVPR, Salt Lake City, UT, USA, 18–22 June 2018.
30. Kim, J.; Jun, J.; Zhang, B. Bilinear Attention Networks. In Proceedings of the 2018 Advances in Neural Information Processing Systems (NeurIPS), Montreal, QC, USA, 3–8 December 2018.
31. Jiang, Y.; Natarajan, V.; Chen, X.; Rohrbach, M.; Batra, D.; Parikh, D. Pythia v0.1: The Winning Entry to the VQA Challenge 2018. *arXiv* **2018**, arXiv:1807.09956.
32. Cadène, R.; Ben-younes, H.; Cord, M.; Thome, N. MUREL: Multimodal Relational Reasoning for Visual Question Answering. In Proceedings of the CVPR, Long Beach, CA, USA, 15–19 June 2019.
33. Cadène, R.; Dancette, C.; Ben-younes, H.; Cord, M.; Parikh, D. RUBi: Reducing Unimodal Biases for Visual Question Answering. In Proceedings of the NeurIPS, Vancouver, BC, Canada, 8–14 December 2019.
34. Clark, C.; Yatskar, M.; Zettlemoyer, L. Don't Take the Easy Way Out: Ensemble Based Methods for Avoiding Known Dataset Biases. In Proceedings of the EMNLP/IJCNLP, Hong Kong, China, 3–7 November 2019.

35. Huang, Z.; Zeng, Z.; Huang, Y.; Liu, B.; Fu, D.; Fu, J. Seeing Out of the Box: End-to-End Pre-Training for Vision-Language Representation Learning. In Proceedings of the CVPR, Nashville, TN, USA, 20–25 June 2021.
36. Rotstein, N.; Bensaïd, D.; Brody, S.; Ganz, R.; Kimmel, R. Fusecap: Leveraging large language models for enriched fused image captions. In Proceedings of the WACV, Waikoloa, HI, USA, 3–8 January 2024.
37. Wang, Z.; Chen, C.; Li, P.; Liu, Y. Filling the image information gap for vqa: Prompting large language models to proactively ask questions. In Proceedings of the EMNLP, Singapore, 6–10 December 2023.
38. Wu, J.; Hu, Z.; Mooney, R.J. Generating question relevant captions to aid visual question answering. In Proceedings of the ACL, Austin, TX, USA, 4–13 October 2019.
39. Banerjee, P.; Gokhale, T.; Yang, Y.; Baral, C. WeaQA: Weak supervision via captions for visual question answering. In Proceedings of the ACL/IJCNLP (Findings), Online, 1–6 August 2020.
40. Zhu, D.; Chen, J.; Haydarov, K.; Shen, X.; Zhang, W.; Elhoseiny, M. Chatgpt asks, blip-2 answers: Automatic questioning towards enriched visual descriptions. *arXiv* **2023**, arXiv:2303.06594.
41. Liu, H.; Li, C.; Wu, Q.; Lee, Y.J. Visual instruction tuning. In Proceedings of the NeurIPS, Vancouver, BC, Canada,9–15 December 2024.
42. Yang, S.; Xiao, L.; Wu, X.; Xu, J.; Wang, L.; He, L. Simple contrastive learning in a self-supervised manner for robust visual question answering. *Comput. Vis. Image Underst.* **2024**, *241*, 103976. [CrossRef]
43. Xiao, L.; Wu, X.; Xu, J.; Li, W.; Jin, C.; He, L. Atlantis: Aesthetic-oriented multiple granularities fusion network for joint multimodal aspect-based sentiment analysis. *Inf. Fusion* **2024**, *106*, 102304. [CrossRef]
44. Shi, X.; Lee, S. Benchmarking out-of-distribution detection in visual question answering. In Proceedings of the WACV, Waikoloa, HI, USA, 3–8 January 2024.
45. Ott, M.; Edunov, S.; Baevski, A.; Fan, A.; Gross, S.; Ng, N.; Grangier, D.; Auli, M. fairseq: A Fast, Extensible Toolkit for Sequence Modeling. In Proceedings of the NAACL-HLT (Demonstrations), Minneapolis, Minnesota, 2–7 June 2019.
46. Socher, R.; Bauer, J.; Manning, C.D.; Ng, A.Y. Parsing with Compositional Vector Grammars. In Proceedings of the ACL, Sofia, Bulgaria, 4–9 August 2013.
47. Hu, M.; Liu, B. Mining and summarizing customer reviews. In Proceedings of the KDD, Seattle, WA, USA, 22–25 August 2004.
48. Pang, B.; Lee, L. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. In Proceedings of the ACL, Barcelona, Spain, 21–26 July 2004.
49. Li, X.; Roth, D. Learning Question Classifiers. In Proceedings of the COLING, Taipei, Taiwan, 26–30 August 2002.
50. Ganapathibhotla, M.; Liu, B. Mining Opinions in Comparative Sentences. In Proceedings of the COLING, Manchester, UK, 18–22 August 2008.
51. Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C.L.; Parikh, D. VQA: Visual Question Answering. In Proceedings of the ICCV, Santiago, Chile, 7–13 December 2015.
52. Lin, T.; Maire, M.; Belongie, S.J.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In Proceedings of the ECCV, Zurich, Switzerland, 6–12 September 2014.
53. Teney, D.; Abbasnejad, E.; van den Hengel, A. Learning What Makes a Difference from Counterfactual Examples and Gradient Supervision. In Proceedings of the ECCV, Glasgow, UK, 23–28 August 2020.
54. Agarwal, V.; Shetty, R.; Fritz, M. Towards Causal VQA: Revealing and Reducing Spurious Correlations by Invariant and Covariant Semantic Editing. In Proceedings of the CVPR, Seattle, WA, USA, 13–19 June 2020.
55. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.C.; Bengio, Y. Generative Adversarial Nets. In Proceedings of the NIPS, Montreal, QC, Canada, 8–11 December 2014.
56. Miller, G.A. WordNet: A Lexical Database for English. *Commun. ACM* **1995**, *38*, 39–41. [CrossRef]
57. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global Vectors for Word Representation. In Proceedings of the EMNLP, ACL, Doha, Qatar, 25–29 October 2014.
58. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In Proceedings of the ICCV, Venice, Italy, 22–29 October 2017.
59. Ma, E. NLP Augmentation. 2019. Available online: https://github.com/makcedward/nlpaug (accessed on 28 October 2024).
60. Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.G.; Salakhutdinov, R.; Le, Q.V. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In Proceedings of the NeurIPS, Vancouver, BC, Canada, 8–14 December 2019.
61. Huggingface. Huggingface: RoBERTa. 2024. Available online: https://huggingface.co/docs/transformers/model_doc/roberta (accessed on 28 October 2024).
62. Ramachandran, P.; Zoph, B.; Le, Q.V. Swish: A self-gated activation function. *arXiv* **2017**, arXiv:1710.05941.
63. Gordon, J.; Durme, B.V. Reporting bias and knowledge acquisition. In Proceedings of the AKBC@CIKM, San Francisco, CA, USA, 27–28 October 2013.
64. Misra, I.; Zitnick, C.L.; Mitchell, M.; Girshick, R.B. Seeing through the Human Reporting Bias: Visual Classifiers from Noisy Human-Centric Labels. In Proceedings of the CVPR, Las Vegas, NV, USA, 26 June–1 July 2016.
65. Hudson, D.A.; Manning, C.D. Learning by Abstraction: The Neural State Machine. In Proceedings of the NeurIPS, Vancouver, BC, Canada, 8–14 December 2019.

66. Malinowski, M.; Doersch, C.; Santoro, A.; Battaglia, P.W. Learning Visual Question Answering by Bootstrapping Hard Attention. In Proceedings of the ECCV, Munich, Germany, 8–14 September 2018.
67. Li, L.; Gan, Z.; Cheng, Y.; Liu, J. Relation-Aware Graph Attention Network for Visual Question Answering. In Proceedings of the ICCV, Seoul, Republic of Korea, 27 October–2 November 2019.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.